Department of Physics and Astronomy
University of Heidelberg

The mathematical structure of measurements, observables and states on Neural Networks

Master's thesis in Physics,
submitted by Johannes Kleiner
2012

This Master's thesis has been carried out by
Johannes Kleiner, born in Kempten (Allgäu),
at the
Institute for Theoretical Physics, University of Heidelberg &
Institute for Physics, University of Freiburg
under the supervision of
Herrn Prof. Dr. Thomas Filk,
Herrn Dr. Harald Atmanspacher,
Herrn Prof. Dr. Ion-Olimpiu Stamatescu.

The mathematical structure of measurements, observables and states on Neural Networks

This thesis studies the mathematical consequences of a notion of "measurement" on NNs, which consists of both, presentation of an input and observation of the output. Once a definition of NNs is chosen, which is general enough to cover a multitude of different NN-models, observables are constructed which represent measurements abstractly. They are found to behave (in light of the fact of NNs being classical systems) unexpected, which is a consequence of this choice of measurement. The properties of the observables are studied. E.g., it is evaluated whether they form a C^*-algebra, which is expected from both classical and non-classical physical systems. In a second part, the relation between the notion of state, which is generated by this concept of measurement, and several consecutive measurements is investigated. Mathematical concepts are proposed which relate the two. This might ultimately amend mathematical tools in experimental situations.

Die mathematische Struktur von Messungen, Observablen und Zuständen auf Neuronalen Netzen

Neuronale Netze sind mathematische Objekte, die in vielen Bereichen der Wissenschaft auftreten, z.B. in der Biologie, Neurowissenschaft, Psychologie, Soziologie und Ökonomie. In all diesen Disziplinen beschäftigt sich die Forschung im Zusammenhang mit Neuronalen Netzen unter anderem damit, die Eigenschaften und das Verhalten von auftretenden Neuronalen Netzen zu studieren. Zu diesem Zweck werden Messungen durchgeführt, die den Zusammenhang zwischen 'Input', der dem Neuronalen Netz präsentiert wird, und 'Output' des Neuronalen Netzes untersuchen.

Diese Arbeit beschäftigt sich mit der theoretischen (mathematischen) Analyse solcher Messungen auf Neuronalen Netzen. Im ersten Teil werden Observablen entwickelt, welche Messungen an einem Neuronalen Netz algebraisch darstellen. Im Weiteren werden ihre allgemeinen Eigenschaften untersucht. Insbesondere wird die Frage studiert, ob die Observablen eines Neuronalen Netzes eine C^*-Algebra bilden, wie dies sonst bei physikalischen Systemen der Fall ist.

Im zweiten Teil der Arbeit werden mathematische Konzepte entwickelt, die erlauben, einen durch obigen Messbegriff induzierten Zustandsbegriff durch Folgen von Messungen zu identifizieren. Insbesondere ermöglichen diese Konzepte, das Verhalten von Neuronalen Netzen trotz eingeschränkter Messmöglichkeiten ohne Nutzung von statistischen Methoden zu studieren.

Acknowledgments

(in German)

Ich möchte mich bei meinen drei Betreuern bedanken. Zum einen bei Harald Atmanspacher und Thomas Filk, welche diese Arbeit in Freiburg betreuten. Ich habe von ihnen viel gelernt, viel mehr als in diese Masterarbeit Eingang finden konnte. Zum anderen bei Ion-Olimpiu Stamatescu, dem ich die Möglichkeit dieser Masterarbeit verdanke, da er sich bereit erklärte, sie in Heidelberg entgegenzunehmen.

Des weiteren geht mein Dank an viele meiner Freunde. Genannt seien Andreas Finke, Clemens Meyer und David Dasenbrock, deren wissenschaftliche Diskussionsbereitschaft mich zu vielen Einsichten brachte, stellvertretend für viele weitere in Regensburg, Heidelberg und Freiburg, denen ich tief verbunden bin.

Dem Institut für Grenzgebiete der Psychologie und Psychohygiene danke ich für die freundliche Bereitstellung eines Arbeitsplatzes und für die vielen interessanten Kontakte, die sich daraus ergaben. Insbesondere danke ich meinen beiden Büro-Nachbarn Rike Wörner und Christian Scheer für die angenehme und freundliche Gesellschaft.

Ohne die Unterstützung meiner Eltern wären diese lehrreiche Zeit in Freiburg sowie mein ganzes Studium nicht möglich gewesen. Vielen Dank dafür! Außerdem bedanke ich mich bei meiner Freundin, Barbara Wittmann, für die Unterstützung und die schöne Zeit, die sie mir bereitete. Last but not least geht mein Dank an meine WG, insbesondere meinen kleinsten Mitbewohner Hannes.

Contents

Introduction

Since McCullouch and Pitts [MP43] first defined what is today called a Neural Network (NN), those mathematical objects have found a widespread application in science [Hay98], ranging from biology and neuroscience [Ami89, DA01, GK02] to psychology [Gar10], sociology [GS09, Gar98] and economics [RZ94]. In all of those fields, NNs are used to describe a multitude of different processes, and research is concerned with the study of the behavior and of properties of the NNs associated with those processes.

Within such research, experiments are carried out which show two generic features: First, the possibility to obtain information about a NN under investigation is limited. Second, the input-dependent behavior of NNs is investigated by studying the relation between the input of the NN and its output.[1]

Thus, a *measurement* consists of presenting an input to the NN under investigation and registering the respective output, which is a function of the states of the neurons of the NN.[2] The scope of this thesis consists of studying measurements, understood in this way, in a theoretical, mathematical way. No application to a specific area of science is presupposed.

Observables are mathematical objects which represent a measurement-process abstractly. E.g., in classical physical theories, observables are typically given by functions on a state-space of a system, and in quantum theory they are given by operators on a Hilbert space. Since they represent measurements, observables and their properties are dependent on the notion of measurement which is presupposed.

Since NNs, as relevant in this work, are classical in nature but subject to measurements which change their state (i.e., invasive measurements), the question arises of which properties to expect of observables associated with those measurements. The non-trivial question of how to construct observables for a general class of NNs constitutes the goal of a first part of this work.

Once observables have been constructed, their properties will be studied. E.g., non-commuting observables arise, which contrasts with classical physical systems, where observables commute.

[1] To illustrate this point, take, e.g., neuroscience. The possibility to obtain information is limited since there exists no measurement technique which allows to read out the states of all neurons of a medium-sized biological organism at once [SK11]. The output could be given by a measurement device, e.g. by the voltage of EEG-electrodes or by the data obtained from an fMRI-scanner.

[2] Note that we use the term "neuron" to refer to systems which are represented by nodes of a NN. No reference to biological neurons shall be implied per se.

This is a direct consequence of associating a process with measurements.[3]

In general, it is an open question of what properties the observables of a system satisfy which is classical in nature but subject to invasive measurements. E.g., observables of classical and quantum-theoretical systems differ in that the latter may not commute, whereas the former always do [Pri90]. However, in both cases they form a C^*-algebra [Mat98, Fil05]. This generates the question of whether this can be expected for all physical systems, in particular if invasive measurements are presupposed. Since NNs as relevant in this work constitute a prime example of invasive measurements, we will investigate whether their observables form a C^*-algebra.

The analysis of the properties of observables of NNs is also interesting from a second point of view, namely when asking about which mathematical structure axiomatically describes a multitude of different (physical and non-physical) systems. There are several approaches which try to answer this question by proposing a mathematical apparatus more similar to the one known from quantum theory [DLM08, FR11]. With respect to NNs, this generates the question of whether a treatment as proposed by such approaches is in principle feasible and of whether the axioms, which are proposed, hold. To answer those questions, we will introduce one of those approaches, GQT, in detail in Sec. 2.2 and evaluate whether it is compatible with the properties of observables on NNs.

Thus, summarized, a first part of this thesis studies the mathematical consequences of the notion of 'measurement' underlying NNs by investigating the structure of observables of NNs.

The second part studies the relation between attractors of a NN and measurements. First, mathematical concepts are developed which allow to identify the attractor of a NN trough series of consecutively performed measurements. In order to do so, some information about the NN is necessary, given by a particular graph. Second, an algorithm is offered which allows to extract this information from a NN through measurements.

It will be shown that in light of the notion of measurement, which is presupposed in this thesis, attractors form a sensible notion of (epistemic) 'states' on NN. Thus, the second part of this thesis contributes to the the question of how a sensible notion of state is mathematically related to measurements as defined above. Therefore, the ideas underlying the proposed concepts might have some relevance for experimental situations. Particularly, in light of the invasiveness of measurements and the limited possibility to obtain information about a NN (c.f. above), they contribute to the question of how a state of a NN can be inferred from what an experimenter observes. Particularly, no statistical averaging is implied. However, it has to be stressed that all results only hold for NNs which are compatible with the definitions of Ch. 1.

The definitions of Ch. 1 do not enforce a certain type of NN but offer a general framework which comprises a multitude of NNs. The definitions are chosen such that they are a) general enough to encompass as many types of NNs as possible but b) specific enough to make a mathematical treatment feasible. There are four conditions which a NN has to fulfill to be compatible with the

[3] Whereas observables of classical systems commute, non-commuting processes are abundant in classical physics.

definitions of Ch. 1:

> It consists of a finite number of neurons.
> Each neuron can only occupy finitely many different states.
> The system can be described in terms of discrete time-steps.
> The update-rule, which specifies the functioning of each neuron (i.e., how it behaves depending on its input), is deterministic (i.e., it does not contain stochastic elements).

Thus the results of this work (concerning observables of NNs on the one hand and mathematical tools to relate a suitable notion of state with measurements on the other hand) hold for all NNs compatible with those constraints.

Concerning the distinction between biological NNs and artificial NNs, where the former term refers to networks of biological neurons in a nervous system, whereas the latter refers to networks of artificial neurons programmed on a computer or directly on hardware, the following can be said: Both terms associate a mathematical model with a real system, be it artificial or biological. Whereas in the artificial case, the model exactly describes the system (because the system has been designed for a given model), in the biological case there usually is some discrepancy between the model and the system (because the model has been designed to mimic a given system). The results of this thesis hold whenever a) a NN - i.e. a real system - is investigated in measurements as described above and b) the model which is associated with the NN obeys the just-mentioned constraints. Particularly, those constraints do not demand the NN-model associated with the real system to be completely specified. E.g., if a NN is known to consist of finitely many neurons which obey the above constraints, the results of this work do hold, irrespective of other properties of the NN such as its topology. However, it may well be that there are practically no biological NNs which are compatible with the above constraints.

This work does *not* aim to contribute to questions about the explicit functioning of biological NNs (neurophysiology) or research concerning the modeling of NNs on a computer.

The structure of this work can be summarized as follows. Ch. 1 offers the definitions relevant for this work and at the same time serves as an introduction into the topic. Ch. 2 is devoted to the development and study of observables of NNs. Ch. 3 develops mathematical tools which relate a suitable notion of 'states' on NNs with series of consecutively performed measurements. Ch. 4 illustrates the results of Ch. 3 and makes some connections with possible applications thereof. Ch. 5 introduces Bell-type inequalities and studies them with respect to NNs. (The reasons for this particular investigation are explained within the chapter.)

Summaries of the results of this work can be found in summary-boxes (orange) throughout this thesis and in the last chapter (p. 87).

Chapter 1

Definitions

This chapter contains definitions and concepts, which are necessary for this thesis. In Sec. 1.1 Neural Networks (NN) are defined. To this end, we will first introduce a graph whose vertices represent neurons and whose edges represent connections between neurons. Neurons may take different states. Therefore, we will introduce a function on the set of vertices which associates a state with each vertex. Similarly, a function will be defined which associates a weight with every edge of the graph.

The dynamics of NNs arise from rules which specify how each neuron's state changes at time t dependent on the states of other neurons at previous times, where the latter are neurons which are connected to the neuron in question by an edge of the graph. Instead of assuming one specific rule, we will chose a general definition.

In order to capture the relation between the dynamics of the NN and its output, which is observed in measurements, a function will be introduced which maps the states of the neurons to an output-set. As demonstrated below, this ensures that the definitions are closer to realistic situations, where the output of a neuron may be given by the output of a measurement-device rather than by neuron-states directly.

The definitions of Sec. 1.1 impose some constraints on NNs, thus not all NNs are compatible with the definitions. However, those constraints ensure a very convenient behavior of attractors, which will be studied in Sec. 1.1.3. A set of attractors will be introduced which is closed in the sense that if an input is presented to the NN once it has settled in one of the attractors, it will consecutively reach another attractor of the set. This set can be used to define a relation which characterizes the behavior of the NN, which will be called attractor-correspondence. It serves as a basis for results of the following chapters.

In Sec. 1.2, the notion of measurement, which is relevant in this thesis, will be defined. Namely, a measurement consists of showing an input to a NN and registering the output which the NN displays. We assume that the latter is evaluated once the NN has settled into an attractor.

In Sec. 1.3, learning-processes are reviewed. This allows to introduce the concept of a perfect

learner, which will be of use in Ch. 2.

A hands-on motivation of why the definitions of Ch. 1 are chosen as outlined can be found in example 1.3 on p. 10.

1.1 Neural Networks

1.1.1 General definition

In the following, a review of the main properties of Neural Networks (NNs) is given. To this end, we introduce definitions which will be used later in this work. The considerations of this chapter are based on, though not taken from, [Hay98, AF06]. Further details can be found in [Hay08, GS09, Kni07, Ami89, GK02, By03, vHS94].

A directed graph $\mathcal{M}$, short digraph, consists of a *finite* set M of vertices, $M = \{v_1, v_2, ..., v_N\}$, together with a 2-valued relation R among this set, i.e. a subset E of $M \times M$, called the edges of the graph [BS03, Wil96, Har11]. We write

$$\mathcal{M} = (M, E). \tag{1.1}$$

$\mathcal{M}$ is called undirected graph (or sometimes simple graph) if the relation R is symmetric, i.e. if $(x, y) \Leftrightarrow (y, x)$, and if elements of $M \times M$ of the form (x, x), called loops, are excluded from E.

Digraphs can be depicted in a graphical form where each vertex is represented by a circle and each edge is represented by a line or an arrow, depending on whether R is symmetric or not. Let an arrow representing an edge of a digraph be pointing towards the latter entry of (v_i, v_j). We will refer to edges (v_i, v_j) of $\mathcal{M}$ by e_{ij}.

In order to obtain a NN, vertices of $\mathcal{M}$ will represent neurons and edges of $\mathcal{M}$ will represent connections between neurons. We associate a state to each neuron at each time. To this end, functions

$$u_t : M \to I \tag{1.2}$$

are introduced, where I is a *finite* representation of the possible states of a neuron, e.g. $I = \{0, 1, 2, 3, ..., I_{max}\}$. In the following, we assume that t is a *discrete* time index.

Additionally, we introduce weights on the edges of $\mathcal{M}$. This mimics, e.g., the variability in synaptic transmission in biological NNs and allows for learning-algorithms to be implemented. Let the function

$$w_t : E \to W \tag{1.3}$$

associate a weight of the possible-weight-set W to each edge of $\mathcal{M}$ for each time-index t. In situations where the time-dependency is clear (e.g. if there is no change in the weights), we abbreviate $w_t(e_{ij})$ as w_{ij}.

The configuration of a NN at time t, $(u_t(v_1), ..., u_t(v_N))$, arises from the configurations at previous times t through algorithmically specifiable rules called **update rules**. I.e., at time step t, the state of every neuron changes according to the update rule and depending on the states and weights of previous time steps, so to speak in parallel. Usually, those rules are local, i.e. $u_t(v_j)$ depends only on those $u_{t-1}(v_i)$ for which $e_{ij} \in E$ as well as the weights of the latter. The following example illustrates the notion of an update rule.

Example 1.1: A typical update-rule

In this example, we illustrate the notion of an 'update rule'. Let the set of possible states of neurons, I, and the set of possible weights of the edges, W, be subsets of the natural numbers, i.e.

$$I, W \subset \mathbb{N}.$$

A typical update rule, which specifies how a state of a neuron changes from one time step t to the next is

$$u_t(v_i) = f\left(\sum_{j:j\to i} w_{ji} u_{t-1}(v_j)\right),$$

where $f : \mathbb{N} \to I$ is a **transfer function** which satisfies

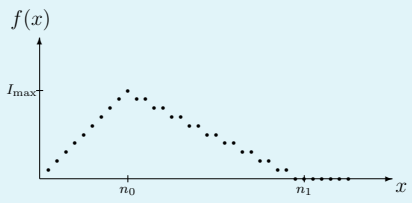
$$\text{Range } f = I,$$

and $j : j \to i$ implies a summation over all j which are connected to i by a directed arrow, i.e. over all $j \in \{j' | e_{j'i} \in E\}$.

The transfer function can, of course, take a variety of forms. Often used are monotonically increasing functions, such as sigmoid functions or step functions [Hay98, Ami89, DA01], but also different forms can be found in the literature. E.g., in [AF06], the following transfer-function is used:

$$f(x) = \begin{cases} \text{Int}\left(I_{max} \cdot (x/n_0)\right) & \text{for } x < n_0 \\ \text{Int}\left(I_{max} \cdot (n_1 - x)/(n_1 - n_0)\right) & \text{for } n_0 \le x < n_1 \\ 0 & \text{for } x \ge n_1, \end{cases} \tag{1.4}$$

where $I_{max} := \max(I)$, $\text{Int}(x)$ rounds x to integers and $n_0, n_1 \in \mathbb{N}$, i.e.:



Throughout this work it is most crucial that we assume a **deterministic update rule**, i.e. an update rule which has no stochastic elements. We do not assume that the update-rule's dependence on previous time-steps has a particular length, but it does have to be finite. Nevertheless, most of the update-rules, which are used in NN-models, only depend on the previous time-step.

1.1.2 Input and output

Conventionally, M is separated into three disjoint sets of vertices, $M_{\mathcal{I}}$, $M_{\mathcal{H}}$ and $M_{\mathcal{O}}$ referring to input-, hidden- and output-vertices, respectively. However, for reasons that will become clear shortly, we will only separate M into two disjoint classes, $M_{\mathcal{I}}$ and $M_{\mathcal{R}}$, i.e.

$$M = M_{\mathcal{I}} \cup M_{\mathcal{R}}. \tag{1.5}$$

Let $M_{\mathcal{I}} = \{v_1, ..., v_{N_{\mathcal{I}}}\}$ and $M_{\mathcal{R}} = \{v_{N_{\mathcal{I}}+1}, ..., v_N\}$.

The **input -neurons** $M_{\mathcal{I}}$ are used to present data to the NN, i.e. they do not actively participate in the dynamics specified by the update-rule. Their states are determined by the data to be presented to the network. To capture this fact mathematically, we define an **input** to be a function which specifies the state of every input-neuron, i.e. every input is a function

$$i : M_{\mathcal{I}} \to I. \tag{1.6}$$

The set of all possible inputs which can be presented to a particular NN is $\mathcal{I}^* = \{i | i : M_{\mathcal{I}} \to I\}$. E.g., if $|M_{\mathcal{I}}| = 1$ (where $|M|$ denotes the cardinality of a set M) and $I = \{0, 1\}$, $\mathcal{I}^*$ consists of two functions: $\mathcal{I}^* = \{a : a(v_1) = 1,\ b : b(v_1) = 0\}$. Usually, only a subset of $\mathcal{I}^*$ is relevant in a given task, which we denote by $\mathcal{I}$.

In the following, we will often use the term 'input i is presented at t_0'. It shall have the following meaning: At t_0, the states of the input-neurons are specified as given by the input i and are kept constant at those states in the following time-steps. At a later time t_1, a different input may be given to the NN, thus, the term 'input i is presented at t_0' is mathematically expressed by

$$u_t(v_j) \equiv i(v_j) \quad \forall\, v_j \in \{v_1, ..., v_{N_{\mathcal{I}}}\}\ \forall\, t :\ t_1 > t \geq t_0, \tag{1.7}$$

We do not allow the time-interval between t_0 and t_1 to be arbitrarily small. Rather, since in Sec. 1.1.3, we will specify the output to be evaluated only when the NN has settled in an attractor, and since measurement consists of presenting an input and registering the output, we allow a different input to be presented only at a time t_1 which is sufficiently distant from t_0, c.f. Sec. 1.1.3. In cases where the time-reference is clear, we will drop the part 'at t_0'. Input-neurons are sometimes called "clamped".

$M_{\mathcal{R}}$ refers to the 'rest' of the vertices, i.e. to all neurons which are not input-neurons. Those vertices are of main interest particularly because they show non-trivial dynamics with respect to the update-rule. Therefore, we call the tuple which consists of the states of the neurons in $\mathcal{M}_{\mathcal{R}}$ at time t the **activation** of the NN, which we denote by $u(t)$. I.e.:

$$u(t) := (u_t(v_{N_{\mathcal{I}}+1}), ..., u_t(v_N)). \tag{1.8}$$

For later convenience, we define $N_{\mathcal{R}} := |M_{\mathcal{R}}| = N - N_{\mathcal{I}}$.

Every activation, which a NN can occupy, determines an output of the NN. Conventionally, the output is defined to be a set of vertices of the NN, particularly a subset of $M_{\mathcal{R}}$. However, we chose a more general definition: We define an output of the NN to be given by a *function* of the states of the neurons. We call the function the **output-function** of the NN and denote it by g_0. Since each activation determines an output, the domain of g_0 is given by all possible activations of a NN, i.e. by $I^{N_{\mathcal{R}}}$, where I is the set of possible states of each neuron. The set to which g_0 maps, i.e. its image-set, is determined by a given situation. E.g., in biological experiments such as EEG-measurements, g_0 may map to tuples of real numbers, as shown in example 1.2. In order to preserve maximal generality, we do not chose a specific image-set, but rather demand g_0 to map to a suitable **output-set** $\mathcal{G}_0$.

Therefore, the output-function g_0 is given by

$$\begin{aligned} g_0 : I^{N_{\mathcal{R}}} &\to \mathcal{G}_0 \\ u(t) &\mapsto g_0(u(t)). \end{aligned} \tag{1.9}$$

The output-function g_0 explicitly captures the relation between the NNs activity and its output, regardless of whether it is given by a measurement device or by the states of a set of neurons. The latter, which is the conventional description, as mentioned above, is contained in the definition of g_0. To see this, suppose that a subset $M_{\mathcal{O}}$ of the neurons of a NN is taken to represent the output of the NN, i.e., the states of the neurons in $M_{\mathcal{O}}$ are the output of the NN. If g_0 is defined to map every activation of the NN to the tuple of states of neurons which are in $M_{\mathcal{O}}$, i.e. if g_0 is the restriction of the activation on the set of neurons in $M_{\mathcal{O}}$, the conventional choice involving explicitly defined output-neurons is recovered.

Thus, a definition of the output of a NN based on an output-function is *more general* than the conventional definition.

We emphasize that g_0 will usually not depend on the states of all neurons of $M_{\mathcal{R}}$. In the following, the term **output-neurons** shall refer to those neurons of $M_{\mathcal{R}}$ whose state may influence the output of the NN defined by g_0.

Fig. 1.1 illustrates the definitions of input and output schematically. Example 1.2 demonstrates that the output-function captures EEG and ECoG-scenarios.

In example 1.3, a hands-on motivation of the definitions of Sec. 1.1.1 and 1.1.2 is given.

Example 1.2: EEG- or ECoG-scenarios

This example serves to illustrate that a definition of NNs in terms of an output-function g_0 is sensible if NNs are studied in EEG- and ECoG-type situations.

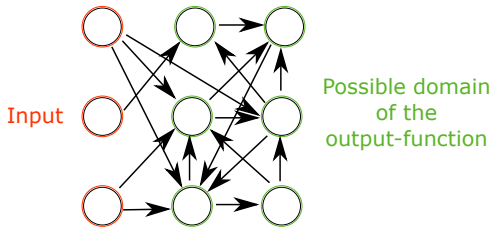


Fig. 1.1: Schematic illustration of the definition of input and output used in this thesis. Note that the output-function can, but may not depend on all non-input neurons.

In both Electroencephalography (EEG) and Electrocorticography (ECoG), several electrodes are used to record electric potentials. The electrodes do not measure the activation of single neurons but an average of the potentials of many neurons. This suggests the following definitions:

For m electrodes let $M_1, ..., M_m \subseteq M_{\mathcal{R}}$ be sets of neurons close enough to individual electrodes for their signal to be picked up by the respective electrode. If we define g_0 to map each activation of the NN to a tuple of m numbers, where the i^{th} number is defined as the average over the states of the neurons in the set M_i, this function covers (in a first approximation) the relation between the states of the neurons and measurement-output.

Of course, more details can be included, e.g. a distance of each neuron to the electrode, or the exact nature of the physical signal which is observed by EEG/ECoG-electrodes.

Example 1.3: A hands-on motivation of the definitions given so far

This work aims at constructing (and proving) mathematical statements about measurements on NNs.

To allow for mathematical statements to be constructed and proofs to be formulated, a set of definitions concerning NNs is required. Since all of the results of this work only hold for NNs compatible with those definitions, this set of definitions should be general enough to encompass a multitude of different NNs. E.g., it should not only allow spiking NNs or only be valid for sigmoid transfer functions. On the other hand, some assumptions have to be made in order to allow for something to be proven at all.

The definitions of Sec. 1.1.1 and 1.1.2 are chosen such that they carefully weigh those two sides. In this example, we will further motivate the most important ones.

Suppose you were yourself about to construct a mathematical statement about NNs. You want the statement to be as general as possible, i.e. to hold for as many types of NNs as possible. How would you proceed?

- **The sets I and W**
  Suppose, your statement should be valid for both spiking NNs without weights on the edges as well as rate-coded NNs with weights taken from real numbers. Instead of treating them case-wise, you can just define: There is a set of possible states called I, and there is a set of possible weights called W. This guarantees that anything proven based on this definition holds for a multitude of different NNs.

- **The update rule**
  In the same manner, suppose your statement should hold for NNs regardless of their transfer function, indeed regardless of whether this is a function at all or not. A sensible way to capture this is to say: There is an update rule which specifies, for each neuron, how the state of the neuron changes depending on the input of the neuron at previous times.

- **The output-function**
  Since the notion of measurement, which is presupposed in this thesis, is motivated by experiments with NNs, the relation between the NN's activity and what an experimenter observes has to be taken into account. If a definition of the form 'out of the N neurons, k serve as the NN's output' were chosen, this would not be compatible with many experimental situations. E.g. in MRI-studies, the activation of the NN is related to what an experimenter observes in a complicated fashion [Sch98]. Therefore, we have settled for the definition: There is a function g_0 which captures the relation between the NN's activity and what an experimenter sees.

It is important to note that we do *not* assume, unless otherwise noted, that explicit knowledge about W, I, g_0 or the update-rule of the NN under investigation is available.

This illustrates the motivation behind the different definitions given so far. As explained above, four constraints had to be added to the definitions. They are summarized in the introduction and in summary 1 (p. 19).

A NN is called **recurrent** if there are edges in $\mathcal{M}$ which form a directed cycle. This is the case if there exists at least one vertex $v_k \in \mathcal{M}$, which allows to walk along the edges of $\mathcal{M}$ (starting at

v_k) such that after some time, v_k is visited again. They form an important class of NN because they allow for dynamics to arise which are non-trivial. E.g., recurrent NNs may have cyclic attractors (c.f. below).

The (approximate) opposite of recurrent NNs are feedforward NNs, which do not contain directed cycles (among other things). This suffices to guarantee a very simple behavior of the NN, which is the reason why we use them to illustrate some results of this thesis in Sec. 3.4.

In the following section, we will study the concept of attractors on NNs. Since we do not want to exclude recurrent NNs from this thesis, a big part of the considerations will deal with cyclic attractors.

1.1.3 Attractors

One of the reasons why NNs are used in a wide range of tasks is their ability - if well trained - to perform complex operations, as e.g. pattern recognition [Hay98]. Since it is the input-output relation which is relevant, the time at which the output is evaluated is of crucial importance.

It is, e.g., not sensible to evaluate the output of the NN at the same time or shortly after the input is presented since a) information may propagate slowly through the NN (this depends on the topology of $\mathcal{M}$ and the update-rule) and b) due to feedback loops (closed paths within $\mathcal{M}$) complex computations may take time to evolve.

Thus, there are two possibilities of when to treat the output of the NN as the *outcome* of an operation. We could either specify a fixed but comparably large time-delay between showing an input and retrieving the output, or specify the output to be evaluated when the NN has reached an attractor. Both are, of course, tantamount to a certain extent, particularly if the fixed time-delay is so long that the NN has settled into an attractor with certainty. Nevertheless, in this thesis, we choose the second option: The output of a NN shall be evaluated only when the NN has reached an attractor. This is especially sensible since an attractor marks the inevitable end of any operation performed by the NN.

In the following paragraphs we will discuss important properties of NNs with respect to attractors. We assume that the weights do not change between presenting an input and retrieving the output.

As specified (p. 6 f.), we deal with NNs which have a deterministic update-rule as well as a finite number of neurons $|M|$ and a finite number of possible states $|I|$. Concerning **attractors** of the NN, this implies the following.

Suppose for the moment that the deterministic update-rule only depends on one past time-step. I.e., the states of the neurons at time t are only influenced by the states of the neurons at time $t-1$. We will generalize this to cases where the update-rule depends on more than one previous time-step below.

As $|M|, |I| < \infty$, there is a finite number of possible configurations of a NN. Since the update rule is deterministic (and due to the assumption of the last paragraph), there is only one trajectory leaving each point in the configuration-space. If a point is visited twice by the same trajectory, the activity has reached an attractor. Thus,

the NN always settles into an attractor in a finite time after an input has been presented. (1.10)

For the same reason,

all attractors are only fixed-point-attractors and limit cycles, (1.11)

other attractor-types do not occur.[1]

The more general case for arbitrary (finite) time-dependence of the update-rule follows in a similar way. Suppose, the update-rules of all neurons depend at most on neuron-states in the k previous time-steps. This implies that the state of each neuron is completely determined by the states of neurons of the last k time-steps. Since the number of configurations is finite, there is also only a finite number of k-tuples of configurations. Thus, after a finite time, the NN has to visit a sequence of k configurations consecutively which it has visited before. Since the update rule is deterministic, this implies that from this point onwards, the trajectory of the NN is exactly the same as before when the NN ran through those k configurations. Thus, the NN has settled in an attractor which is either a fixed-point attractor or a limit cycle.

To maintain a certain degree of simplicity in the following considerations, we will stick with the assumption of a 1-time-step dependence of the update rule. However, a generalization analogous to the last paragraph can be applied in all cases.

We will denote an attractor by z. It consists of a finite sequence of activations (p. 8), i.e.,

$$z = (u(t_1), ..., u(t_{n_z})). \tag{1.12}$$

n_z is the length of the cycle, where $n_z = 1$ for fixed-point-attractors and $n_z > 1$ for cyclic attractors. Due to the finite number of activations, n_z is finite, i.e. $1 \leq n_z < \infty$.

This choice is **non-trivial** since we have defined the activation $u(t)$ to only comprise the states of those neurons which are in $M_{\mathcal{R}}$, but not the states of the whole set of neurons (p. 8). This is justified since the input-neurons do not participate in the dynamics. A consequence of this is, however, that a NN may be in the same attractor z even for different inputs. Needless to say, this is not the general case: the attractor, which a NN inhabits, essentially depends on the input which presented to the NN.

[1] This follows from the following consideration. Suppose, the NN follows an arbitrary trajectory in its configuration-space. Due to the finite number of activations, this trajectory has to visit a configuration which it has visited before after a finite time. Since the update-rule of the NN is deterministic (and since we currently assume that its time-dependence is of length 1), from this point onwards the trajectory will follow exactly the path it has taken before. Thus, the NN has settled either in a fixed-point attractor or in a limit cycle.

1.1.3.1 Recognizing attractors

Since this work deals with situations in which only the output of a NN is observable, a clear understanding of the relation between the behavior of an attractor $z = (u(t_1), ..., u(t_{n_z}))$ and the behavior of the output is necessary. In this subsection, we will show that if a NN occupies an attractor, then the output of the NN is periodic, and vice versa: if a periodic output is observed for a certain number of time-steps, then the NN has to have settled in an attractor.

The first of those claims, i.e., that once the NN has settled into an attractor, the output of the NN behaves periodically can easily seen by the following argument. If a NN has settled into an attractor with period n_z at a time t_0, and if we count the time-steps following t_0 by $t_1, t_2, ...$, then

$$\forall\, n > n_z : u(t_n) = u(t_{n-n_z}).$$

Since the output is a function of the activations (i.e., to every activation only one output is associated), it follows that

$$\forall\, n > n_z : g_0(u(t_n)) = g_0(u(t_{n-n_z})).$$

Therefore, the output shows periodic behavior. Particularly, since several activations may have the same output, this shows that the period of the output is smaller than or equal to the period of the attractor.

We now address the question of whether we can tell from the output of a NN that it has settled in an attractor.

There is only a finite number of activations of a NN. Since activations are $|M_{\mathcal{R}}|$-tuples of elements of the possible-state-set I, there are $|I|^{|M_{\mathcal{R}}|}$ possible activations. Furthermore, whenever an activation is visited twice, the NN has reached an attractor (compare Eq. (1.10)). Thus, once $|I|^{|M_{\mathcal{R}}|}$ time-steps have passed after the presentation of an input, the NN has certainly settled into an attractor.

This means that if a NN is allowed to run for $|I|^{|M_{\mathcal{R}}|}$ time-steps, it certainly occupies an attractor. A lower bound can be given, if the activations are considered which have a given output o_i, as given by the set

$$\{u(t) | g_0(u(t)) = o_i\}. \tag{1.13}$$

For a cycle of outputs, say $(o_1, o_2, ..., o_n)$, the activations which could be occupied by the NN at any time during the cycle are given by unions of the sets (1.13). The cardinality of this union is given by

$$N_{(o_1,o_2,...,o_n)} := |\{u(t) | g_0(u(t)) \in \{o_1, o_2, ..., o_n\}\}|. \tag{1.14}$$

Thus, for the same reason as above, once a cyclic output $(o_1, o_2, ..., o_n)$ has been observed for $N_{(o_1,o_2,...,o_n)}$ time-steps, the NN certainly occupies an attractor.

This shows that a NN can always be verified to occupy an attractor by looking at the output. $N_{(o_1,o_2,...,o_n)}$ may not be known in realistic situations and may have to be estimated. Effectively, $N_{(o_1,o_2,...,o_n)}$, or an estimation thereof, specifies the number of time-steps between t_0 and t_1 as defined in the context of Eq. (1.7).

1.1.4 Closed set of attractors

We have mentioned before that a generic feature of measurements on NNs is that they only allow for limited information to be obtained from NNs (p. 1), which implies that individual activations of a NN may in many cases not be distinguishable by observation of the output. This motivates the following considerations, which aim to construct a description of NNs solely based on attractors. We will see later (Ch. 3) that the concepts which describe the behavior of NNs on the basis of attractors can be recovered by measurements on NNs.

The goal of the following considerations is to construct a description of NNs with attractors only. To achieve this, we consider a set of attractors which has the following property: Whenever the NN resides in one attractor, if an input is given to the NN, the NN will consecutively reach an attractor which is also in this set. This will be called a closed set of attractors. All later considerations will use this description, in particular the algebraic treatment of the observables of NNs.

If an input i is presented to a NN which resides in activation $u(t)$, the NN will consecutively run into an attractor z (c.f. Eq. (1.10)). We denote this by

$$(i, u(t)) \dashrightarrow z.$$

We define a set of attractors $\mathcal{Z}$ to be **closed** if the following equation holds. It expresses the fact that for every attractor in the set $\mathcal{Z}$, every input out of a fixed input-set $\mathcal{I}$ (c.f. p. 8) can be shown to the NN at any time-step of the attractor's cycle (given by $(u(t_1), u(t_2), ..., (u_{n_z}))$, c.f. Eq. (1.12)) such that the attractor which the NN reaches subsequently is also in $\mathcal{Z}$:

$$\forall\, z \in \mathcal{Z},\ \forall\, i \in \mathcal{I} : \forall\, k \in \{1, ..., n_z\}\ (i, u(t_k)) \dashrightarrow z' \text{ such that } z' \in \mathcal{Z}. \tag{1.15}$$

Furthermore, we define a closed set of attractors $\mathcal{Z}$ to be **minimally closed** if all $z' \in \mathcal{Z}$ are removed from $\mathcal{Z}$ for which the following holds: There is no attractor in $\mathcal{Z}$ to which an input i can be shown such that the NN subsequently reaches z'. I.e, $\mathcal{Z}$ is minimally closed if all $z' \in \mathcal{Z}$ are removed which satisfy

$$\neg\big(\,\exists\, i \in \mathcal{I},\ \exists\, z \in \mathcal{Z} : \exists\, k \in \{1, ..., n_z\} \text{ with } (i, u(t_k)) \dashrightarrow z'\big) \tag{1.16}$$

($\neg$ denotes the negation of the statement in brackets). We define a set of attractors to be a **closed set of attractors** if it is minimally closed. In the following, $\mathcal{Z}$ always denotes a closed set of attractors. Once one particular attractor and a set of inputs are specified, the closed set of attractors is unique.

A closed set of attractors $\mathcal{Z}$ is **finite**, i.e. has finite cardinality. This is a consequence of the number of neurons and the number of possible states of neurons being finite, from which it follows

that the number of *activations* of a NN is finite. The latter implies that the number of attractors is finite since one activation cannot be part of more than one attractor, c.f. p. 13.

After having introduced a closed set of attractors above, the next step towards achieving a description of a NN in terms of attractors is to expand the definition of the output-function g_0 to map attractors rather than activations of a NN into a set of outputs in a natural way. We chose the output-set of the expanded function to be a cartesian product of output-sets of g_0. This allows us to view the output of cyclic attractors not as (possibly) varying series in $\mathcal{G}_0$, but as a point in the new output-set.

The output-function g_0 has been defined as (Eq. (1.9))

$$g_0 : I^{N_\mathcal{R}} \to \mathcal{G}_0. \tag{1.17}$$

Thus, g_0 can naturally be applied to fixed-point-attractors $z = u(t_1)$ of $\mathcal{Z}$. In order to treat cyclic attractors, we chose n_{max} to be the largest among all periods of the attractors of $\mathcal{Z}$ and define

$$\mathcal{G}_\mathcal{Z} := \underbrace{\mathcal{G}_0 \times .. \times \mathcal{G}_0}_{n_{max}}.$$

Thus, a point in $\mathcal{G}_\mathcal{Z}$ is a series in $\mathcal{G}_0$.

This allows us to define the **generalized output-function** $g_\mathcal{Z}$ as

$$g_\mathcal{Z} := \underbrace{g_0 \times ... \times g_0}_{n_{max}}. \tag{1.18}$$

In order to assure that $g_\mathcal{Z}$ is well-defined also for cyclic attractors which have a period smaller than n_{max}, we define $\emptyset$ to denote the absence of an object and $g_0(\emptyset) := \emptyset$. This implies that for an arbitrary attractor z of with period n_z,

$$g_\mathcal{Z}(z) = g_\mathcal{Z}(u(t_1), ..., u(t_{n_z}), \emptyset, ..., \emptyset) = (g_0(u(t_1)), ..., g_0(u(t_{n_z})), \emptyset, .., \emptyset). \tag{1.19}$$

Therefore, the generalized output-function $g_\mathcal{Z}$ is a well-defined object with

$$g_\mathcal{Z} : \mathcal{Z} \to \mathcal{G}_\mathcal{Z}, \tag{1.20}$$

$$z \mapsto g_\mathcal{Z}(z). \tag{1.21}$$

1.1.5 The attractor-correspondence

So far we have reformulated two important concepts associated with NNs in terms of attractors. First, we have introduced a closed set of attractors, which specifies (in a sense specified in Ch. 3) the state of a NN. Second, we have defined a generalized output-function which specifies the output associated with each attractor.

In the following we will introduce another object which completes the description of the behavior of NNs on the level of attractors. This object, which we call attractor-correspondence, specifies how the attractor of a NN changes once an input is presented to the NN. If an attractor of $\mathcal{Z}$ and an input $i \in \mathcal{I}$ is given, the attractor-correspondence specifies which attractors the NN may reach subsequently. Thus, together with the generalized output-function, it captures the behavior of a NN.

Since cyclic attractors consist of several activations, and since it is the latter rather than the former which determines which attractor the NN runs into once an input is presented, the attractor-correspondence has to allow for more than one attractor to be associated which each pair of input and attractor.

The essential point in the concept of a function which maps elements of a set $\mathcal{A}$ to a set $\mathcal{B}$ is that to each element of $\mathcal{A}$, it only associates one element of $\mathcal{B}$, not more. However, as just explained, the attractor-correspondence has to associate more than one element of $\mathcal{Z}$ to every element of $\mathcal{I} \times \mathcal{Z}$. Hence, it is *not* a function. In the following, we will define the attractor-correspondence, comment about its mathematical nature and introduce probabilities, which are associated with it.

We denote the **attractor-correspondence** by

$$\begin{aligned} f : \mathcal{I} \times \mathcal{Z} &\rightharpoonup \mathcal{Z}, \\ (i, z) &\mapsto f(i, z) \end{aligned} \tag{1.22}$$

where the symbol $\rightharpoonup$ is chosen to denote that it is not a function, i.e. that $f(i, z)$ can be a set of elements of $\mathcal{Z}$.

We define $f(i, z)$ by the following equation. It states that every attractor $z' \in \mathcal{Z}$ is in $f(i, z)$ if and only if there exists at least one activation $u(t)$ contained in attractor z such that the NN settles in attractor z' subsequently to the presentation of input i to this activation:

$$z' \in f(i, z) \Leftrightarrow \exists\, u(t) \in z : (i, u(t)) \dashrightarrow z'. \tag{1.23}$$

Thus, $(i, z, f(i, z))$ can be understood as a subset of (or a relation on) $\mathcal{I} \times \mathcal{Z} \times \mathcal{Z}$. Mathematically, f is called a correspondence.

For later convenience, we define the term 'attractors $f(i, z)$ maps to' denote all attractors $z' \in f(i, z)$.

Note that the attractor-correspondence only associates more than one attractor to every pair (i, z) if z is a *cyclic* attractor. Since fixed-point-attractors z_{fp} have a periodicity of 1, i.e. comprise one activation only, only one attractor is associated with each pair (i, z_{fp}). I.e.,

$$|f(i, z_{fp})| = 1 \ \forall\, i \in \mathcal{I}. \tag{1.24}$$

This implies that f behaves as a function if its second argument is a fixed-point-attractor.

Up to this point, we have taken a general (c.f. p. 3) definition of NNs and constructed a description in terms of attractors based on it. Since attractors of NNs can be cyclic, the description in terms of attractors is not deterministic. Rather, whenever $|f(i,z)| > 1$, the NN may run into several different attractors once input i is presented to attractor z. In the following, we **associate probabilities** in a natural way to those different possibilities.

Let $z_k, z_l \in \mathcal{Z}$, $i_m \in \mathcal{I}$. Define $N_{i_m}(z_l|z_k)$ to be the set of activations contained in attractor z_k for which the following statement holds: If input i_m is presented to the NN while it occupies one of those activations, the NN subsequently settles into attractor z_l. I.e.,

$$N_{i_m}(z_l|z_k) := \{u(t) \,|\, u(t) \in z_k,\ (i_m, u(t)) \dashrightarrow z_l\}. \tag{1.25}$$

If there is no further information about the activations which are contained in an attractor, and particularly, if there is no information about the order, in which the activations are aligned in a cycle of an attractor, the following probabilities specify how likely it is that the NN, if residing in attractor z_k, runs into attractor z_l subsequent to the presentation of input i_m. They are given by the number of activations in a cycle of z_k, which map to z_l upon presentation of input i_m divided through the period n_{z_k} of the attractor z_k, thus by

$$p_{i_m}(z_l|z_k) := \frac{|N_{i_m}(z_l|z_k)|}{n_{z_k}}. \tag{1.26}$$

This is a consequence of the fact that every activation $u(t)$ can only be realized for one time-step in one period of a cyclic attractor, c.f. Eq. (1.10). Since those probabilities specify the behavior of the NN on an epistemic level where individual activations of the NN cannot be discriminated, we refer to them as **epistemic probabilities**.

Observe that Eq. (1.26) equals 1 or 0 if z_k is a fixed-point-attractor (i.e. the behavior of a fixed-point attractor is not probabilistic), and that

$$\sum_{z_l \in \mathcal{Z}} p_{i_m}(z_l|z_k) = 1 \tag{1.27}$$

for all (cyclic or fixed-point) attractors.

1.1.6 Summary

A summary of which NNs are compatible with the definition of Sec. 1.1, and thus with the results of this thesis, can be found in the introduction (p. 3), as well as in the following summary-box. At this point we make a note about which type of NNs are *not* compatible with the definitions.

Out of the four constraints the most restrictive seems to be the finiteness of the number of possible neuron-states, i.e. the finiteness of $|I|$. Whenever a NN (or a model thereof) is given which contains a continuous neuron state variable, the results of this thesis do not hold. The reason for this is that in such a case, there may be an infinite number of activations which the NN can occupy and thus, it cannot be guaranteed that there are only fixed-point or cyclic attractors.

However, a more elaborate treatment of attractors may allow to generalize the results of this work also for systems with continuous state variables. The same is true for all of the other restrictions. Once all kind of attractors (i.e., next to fixed-point and cyclic attractors, also strange attractors and limit tori) can be treated in a way similar to the one of sections 1.1.3, 1.1.4 and 1.1.5, the results of this thesis can most likely be generalized to arbitrary NNs.

However, in anticipation the results of Ch. 2 and 3, we note that in order to make the concepts of this thesis applicable to experiments, further changes might be necessary, e.g., the inclusion of measurement-errors or noise into the considerations.

Summary 1: Defining NNs

Sec. 1.1 defines (and hence introduces) NNs. Since all mathematical statements of the following chapters are based on those definitions, and thus only hold for NNs which are compatible with them, they are chosen to be as general as possible but at the same time specific enough to offer sufficient details for a mathematical treatment to be feasible. Consider example 1.3 (p. 10) for a further elaboration of this point.

In particular, we have defined

- $\mathcal{I}$ to be the set of **inputs** that is presented / presentable to the NN. (p. 8)
- $\mathcal{Z}$ to be the **set of attractors** of the NN which is closed with respect to $\mathcal{I}$. This means that each input of $\mathcal{I}$ can be shown to the NN being in any attractor $z \in \mathcal{Z}$ such that the new attractor, which the NN reaches, in is also in $\mathcal{Z}$. It is also minimally closed in that there is no attractor in $\mathcal{Z}$ which is not reachable from other attractors in $\mathcal{Z}$ by at least one input. (p. 12, 15)
- f to be the **attractor-correspondence**, which is a function for fixed-point-attractors but more general than a function for cyclic attractors: $f : \mathcal{I} \times \mathcal{Z} \rightharpoonup \mathcal{Z}$. It specifies which attractors the NN can reach upon presentation of the input i given the attractor z. Probabilities can be associated with different possibilities, c.f. Eq. (1.25) and (1.26). (p. 16, 18)
- $g_{\mathcal{Z}}$ to be the **generalized output-function**, $g_{\mathcal{Z}} : \mathcal{Z} \to \mathcal{G}_{\mathcal{Z}}$. It specifies which output-value in a suitable output-set $\mathcal{G}_{\mathcal{Z}}$ belongs to a given attractor. In general, $\mathcal{G}_{\mathcal{Z}}$ is a set comprising series of 'direct'-NN-output. (p. 16)

NNs are compatible with the definitions of this section if

they comprise a finite number of neurons, i.e. M is finite. (p. 6)
each neuron may only occupy finitely many different states, i.e. I is finite. (p. 6)
a discrete time-index exists in which the system can be described. (p. 6)
the update-rule, which specifies the functioning of each neuron, is deterministic. (p. 8)

1.2 Measurement

As explained in the introduction, this work focuses on a notion of measurements which consists of presenting an input to the NN under investigation *and* registering its output. This is non-trivial since such a measurement is a process, i.e., it changes the system under investigation.

In this work, we investigate the mathematical structure of measurements in a twofold way. First, we will describe the process of making a measurement mathematically, based on the definitions of a NN as given in Sec. 1.1. All of the concepts necessary to do so have already been constructed, namely the notion of a closed set of attractors, the attractor-correspondence and the generalized output function.

Second, we will construct observables which represent measurements abstractly. Namely, to every possible measurement, we will associate an operator (i.e., a matrix) which encodes information about this measurement. This is common practice in physics since it allows to study properties of measurements in a mathematical way. Concerning NNs, e.g., the construction of observables will allow us to formulate statements about the set of all possible observables of NNs compatible with the definitions of Sec. 1.1. Observables of NNs will be constructed in Ch. 2.

Note that even though observables are associated with every measurement, the latter is something completely different from the former. Whereas a measurement is an explicit specification of how to treat a system in order to obtain specific information about it, an observable is an abstract mathematical object. This is reflected, e.g., in the fact that in quantum theory, if an arbitrary observable is specified (c.f. Sec. 2.1.2), there is no general method which specifies how a measurement corresponding to this observable can be carried out [Fil05].

1.2.1 Epistemic and ontic descriptions

A system can be described in ontic and epistemic terms. Since this distinction is reflected in the scope of this thesis, it will be introduced in this section. We restrict attention to classical systems.

Suppose a system is given and a model thereof is specified, i.e., particularly, the state-space of the system can be constructed. Ontic states are defined to be points of this state space. Thus, they completely specify the configuration of the system.

In contrast, epistemic states are probability distributions (or, to be more precise, probability measures of a probability space) over the state-space of the system, whose support is a subset of the state-space. Thus, epistemic states do not completely specify the configuration of a system. Rather, if a system occupies a particular epistemic state, this expresses that some (but not all) knowledge about the configuration of the system is available.

This difference between ontic and epistemic states defines ontic and epistemic descriptions. If a system is described in terms of ontic states, this is a **ontic description** of the system, and if it is described in terms of epistemic states, this constitutes an **epistemic description** of the

system. A good example for this are NNs in experimental situations. As we have stressed before, there are only limited possibilities to carry out investigations on NNs in experiments, which is captured by the definition of measurement given below. Particularly, the possibility to obtain information about the configuration of the NN is limited, which in turn implies that ontic states cannot be identified.[2]

Thus, in this terminology, part of the scope of this thesis is to develop an epistemic description of NN which is compatible with the presupposed notion of measurement. Since measurements do not only limit the possibility to obtain information from the NN, but are also invasive (state-changing), the mathematical object, on which this description is based, cannot simply be probability distributions over the configurations of a NN.

Rather, the mathematical objects which can legitimately be called 'epistemic states' of a NN with respect to the relevant notion of measurement are the attractors of a NN. This is so because a) the behavior of NNs can be specified based on attractors (we have constructed a description in terms of attractors in Secs. 1.1.4 and 1.1.5), and b) because the attractor which a NN occupies can be identified by measurements, which will be shown in Ch. 3.

Thus, summarized, this work evaluates in how far an epistemic description of NNs in light of the presupposed notion of measurement is feasible. To this end, a description based on attractors is constructed. Observables, which are associated with measurements and thus with this epistemic description, will be developed and studied.

We emphasize that we have introduced the distinction between epistemic and ontic states with respect to a model of a system, not with respect to reality per se. The latter discussion is philosophical in nature and therefore not part of the scope of this thesis.

Further information about this topic can be found in [bGFA12] and [AP05]. Whereas [bGFA12] focuses on classical systems, [AP05] introduces this terminology with respect to quantum theory.

1.2.2 Mathematical description of measurements

This work presupposes a specific notion of measurements on NN: Measurements consist of showing an input to the NN under investigation and obtaining its output. In Sec. 1.1, we have defined the output of a NN to be given by an output-function, which maps the states of the neurons of the NN (i.e., activations of the NN) to an output-set $\mathcal{G}_0$ (p. 9). Furthermore, we have introduced a generalized output-function (p. 16) which maps attractors (i.e. sequences of activation) to an output-set $\mathcal{G}_{\mathcal{Z}} = \mathcal{G}_0 \times ... \times \mathcal{G}_0$.

Since we have defined the output of a NN to be evaluated once the NN has settled into an attractor (c.f. Sec. 1.1.3), a measurement as relevant in this work consists of an input i, which

[2] If NNs are defined as in Sec. 1.1, and if the weights of a NN and the input are fixed, ontic states are given by the activations of the NN (as defined on p. 8). This is so because a specification of the states of all neurons in $M_{\mathcal{R}}$ determines the dynamics of the NN in this case, and hence generates its state-space.

is presented to the NN and an element of $\mathcal{G}_\mathcal{Z}$, which is the output of the NN subsequent to the presentation of input i. I.e., a **measurement** is a pair

$$(i, o) \qquad \text{where } i \in \mathcal{I},\ o \in \mathcal{G}_\mathcal{Z}. \tag{1.28}$$

Throughout this work, we chose o (as "output") to denote an element of $\mathcal{G}_\mathcal{Z}$.

Suppose the NN under investigation settles in attractor $z \in \mathcal{Z}$. If a measurement corresponding to input i is performed, i.e. if input i is presented to the NN, the NN will subsequently settle in an attractor $z' \in f(i, z)$. Thus, the output of the NN is $g_\mathcal{Z}(z')$, and the measurement is given by

$$(i, o') \qquad \text{with} \quad o' \in \{\, g_\mathcal{Z}(z') \mid z' \in f(i, z)\}. \tag{1.29}$$

1.2.3 Histories of measurements

We define a **history** h to be a series of consecutively performed measurements. I.e.,

$$h = \big((i_1, o_1), ..., (i_k, o_k)\big), \tag{1.30}$$

where $i_1, ...i_k \in \mathcal{I}$, $o_1, ..., o_k \in \mathcal{G}_\mathcal{Z}$. We call k the **length** of h or refer to h as a k**-history**.

Summary 2: Measurement

A **measurement** consists of showing an input i and obtaining the respective output o. It is mathematically represented by the pair (i, o), where $o \in \mathcal{G}_\mathcal{Z}$ is obtained once the NN has settled in an attractor (c.f. p. 12). A **history** h is a series of consecutively performed measurements.

1.3 Learning

In the this section, we review the concept of learning (also called training) on NNs. On the one hand, this section serves as quick introduction into the topic. On the other hand, it allows to define an important class of NNs which will be used later in this thesis.

It has already been noted that one of the essential advantages of NNs is their ability to perform complex operations. Of course, the type of operation varies from task to task and the challenge of NN-design lies in the question of how to construct a NN that will optimally perform in a given situation, i.e. it lies in the choice of $\mathcal{M}$, I, W, g_0 and of the update rule. Once a task is defined and the NN-design has been chosen, a learning-process is applied in order to optimize the NN-behavior by subsequent changes of the weights of the edges of the NN.

There are two basic classes of learning processes, supervised learning and unsupervised learning, which will be described in the following.

1.3.1 Supervised learning

The first class is called supervised learning, or 'learning with a teacher'. In this class, a set of data consisting of pairs of input and output is available which determines what the optimal output for a given input is, i.e. a set D with

$$D \subseteq \{(i, o_i') | i \in \mathcal{I}, o_i' \in \mathcal{G}_{\mathcal{Z}}\}. \tag{1.31}$$

The learning-process consists of sequential changes of the weights of the edges of the graph $\mathcal{M}$ underlying the NN (resp. adding/deleting edges). After each change, the performance of the NN is evaluated. To do so, one defines a parameter ε which evaluates how much the output of the NN corresponding to each input differs from the output o_i' defined by the data-set D. Since the performance of the NN, and thus ε is depending on the weights of the NN, ε can be understood as an *error-function* on those weights. It is defined such that ε is 0 if the NN performs exactly as specified in D.

How to define ε suitably is a non-trivial question, for first, it will have to weigh changes in outputs belonging to different inputs, e.g., if the NN's performance w.r.t one input improves while does the opposite for another input. And second, since it may induce order effects. E.g., if the inputs are always shown in a certain order, the NN may only perform optimal with respect to this order of inputs [AF06].

Suppose a parameter ε which evaluates the performance of the NN-behavior within a learning-process has been defined. The graph of ε over the space of weight-configurations defines an "error-surface", on which the NN propagates. Thus, the goal of a learning-process consists of finding a global (or a particularly deep local) minimum of this error-surface.

The question of how to vary the weights of a NN suitably so as to quickly achieve an optimal performance depends on the type of NN as well as the on aimed-at task and is certainly part of the challenge of NN-design. There are many algorithms available, an example of which is given in the following.

Example 1.4: Genetic algorithm

A big class of learning-process is called genetic algorithm. The idea is to change the weights of a NN either arbitrarily or according to some rules. Whenever a change has occurred, the performance of the NN, as given by ε, is evaluated and only if ε has decreased, the change is accepted. If it has not decreased, the weights are put back into the configuration before the last change.

An abundant treatment of supervised learning-processes can be found in [Hay08].

1.3.2 Unsupervised learning

The second class of learning-processes is called unsupervised learning or 'learning without a teacher'. There are two concepts underlying this class.

One corresponds exactly to the supervised learning-class, except that there is no data-set D present. Rather, a direct environmental feedback-signal is available which replaces the error-function ε. Again, changes in the weights are made in order to optimize the NN-behavior with respect to the feedback-signal. If the environmental feedback-signal is boolean, i.e., if it consists of 'yes' and 'no' only, this is called reinforcement learning.

The second class refers to situations where no environmental feedback is present, as e.g. the case with self-organizing maps, c.f. [Hay98, Ch. 9]. NNs belonging to this class completely evolve according to their own dynamics once an input is presented.

Unsupervised learning-tasks are not studied further in this thesis. Further information can be found in [Hay08].

1.3.3 Perfect learner

In many situations which involve supervised NN-learning, the trained NN is required to recognize the inputs on which it was trained independent of the order of presentation. In accordance with [AF06] we adopt the following terminology.

A NN is called a **perfect learner** if and only if it satisfies the two criteria:

- The performance of the NN, as measured by the error-function ε, is optimal, i.e., $\varepsilon = 0$.
- The output of the NN is independent of the attractor in which the NN settles when a given input is shown.[3]

Given the definitions of Sec. 1.1, we can express the latter statement in the following formula. It specifies that the output associated with those attractors, which the NN can reach if input i is presented to an attractor $z \in \mathcal{Z}$, is independent of $z \in \mathcal{Z}$:

$$g_{\mathcal{Z}}(f(i, z_k)) = g_{\mathcal{Z}}(f(i, z_l)) \qquad \begin{array}{l} \forall\, i \in \mathcal{I}, \\ \forall\, z_k, z_l \in \mathcal{Z}, \end{array} \tag{1.32}$$

where we write $g_{\mathcal{Z}}(\{a, b, c\})$ instead $\{g_{\mathcal{Z}}(a), g_{\mathcal{Z}}(b), g_{\mathcal{Z}}(c)\}$ for notational convenience. ($f(i, z_k)$ may consist of more than one attractor.)

Of course, both properties are important, e.g., in industrial applications of NNs. But also in biological scenarios, NNs behaving as a perfect learner can be expected to be found: Evolution may require organisms to correctly recognize prey or predators independent of their mental state beforehand.

[3] Note that this does not follow from the first point. As elaborated above, several choices for ε are possible, many of which may provoke order-dependent behavior.

Perfect learners can be obtained, e.g., by use of genetic algorithms [AF06].

In Sec. 2.4, we will use the class of perfect learners to establish some properties of observables of NNs.

Summary 3: Learning

Sec. 1.3 reviews the notion of a learning-process.

In both supervised and unsupervised learning, changes are applied to the weights of the edges of $\mathcal{M}$ in order to optimize the input-output behavior of a NN. In the former case, a set of data D is available, which specifies the optimal performance of the NN. (p. 23)

If a NN reproduces the input-output-behavior specified in D exactly and independent of the order of presentation of the inputs, we call this NN a **perfect learner**. (p. 24)

Chapter 2

Observables of NNs

The goal of this chapter is to construct observables which represent measurements on NN. Once this is done, we will study some of their properties.

This can be motivated in a twofold way. On one hand, observables of both quantum systems and classical systems form a C^*-algebra, which will be introduced in detail in Sec. 2.1. Thus, the question arises of whether the observables of any system form a C^*-algebra. We will answer this question in Sec. 2.4, where we show that the set of observables of NNs does not form a C^*-algebra.

Therefore, the question arises of which mathematical structure can be assumed to hold for systems in general. Is there a set of axioms which holds for classical and quantum-physical systems as well as for non-physical systems? Research in this direction is being carried out and several results have been reported [DLM08, FR11]. In Sec. 2.2, we will introduce one approach which provides answers to this question. It is called 'Generalized Quantum Theory'. In order to evaluate whether this mathematical framework is general enough, we will first evaluate whether a treatment as proposed by this approach is feasible on NNs, and second evaluate whether its axioms hold for observables of NNs (Sects. 2.3 and 2.4). This constitutes a second motivation to investigate observables on NNs.

2.1 Motivation 1: C^*-algebras

As has been noted above, observables of both classical and quantum-theoretic physical systems obey the structure of a C^*-algebra. One of the foci of this work is to evaluate whether this is true also for the observables of a NN. To this end, we will introduce the concept of a C^*-algebra in Sec. 2.1.1. In Sec. 2.1.2, some explanations are offered as to why C^*-algebras arise in physics.

2.1.1 The path to C^*-algebras

In the following, we introduce the concept of a C^*-algebra. To the benefit of the reader, the presentation has been chosen to be slightly pedagogical in nature.

The following definitions and explanations are taken from [BEH99, Appendix B] and [Fil05, Ch. 3], but valuable introductions can also be found in [Mat98] and [HS85].

What is a complex algebra?

Suppose you are given a set M. A binary operation on M is a function $\phi : M \times M \to M$. We call it *associative* if and only if (iff) $\phi(\phi(a,b),c) = \phi(a,\phi(b,c))$ and *commutative* iff $\phi(a,b) = \phi(b,a)$.

A set M equipped with an *associative* binary operation is called a semi group. If additionally there exists a *unit element* $e \in M$ such that $\phi(g,e) = \phi(e,g) = g$ for all $g \in M$ and an *inverse element* $g^{-1} \in M$ for all $g \in M$ such that $\phi(g,g^{-1}) = \phi(g^{-1},g) = e$, M is called a group.

Suppose we are given a set R with two *associative* binary operations, one called *summation*, ϕ_S, and one called *multiplication*, ϕ_M, for which we write $a + b$ and ab, respectively, i.e. $a + b := \phi_S(a,b)$ and $ab := \phi_M(a,b)$. We call (R, ϕ_S, ϕ_M) a ring if (R, ϕ_S) is a commutative *group* and both operations are *distributive*, i.e. $a(b+c) = ab + ac$ and $(a+b)c = ac + bc$ for $a, b, c \in R$.

Note that in accordance with [BEH99, Fis05], we *do not demand* that R has an unit element with respect to (w.r.t.) the multiplication, i.e. an element $e \in R$ such that $ae = ea = a$ for all $a \in R$. If it does have one, we say R is a ring with unit element.

In order to construct a complex algebra, which is the basis for a C^*-algebra, we need a complex vector space $\mathcal{V}$. It is defined as follows [Fis05].

A complex vector space $\mathcal{V}$ is a set V together with a binary operation called vector addition, $+ : V \times V \to V$, $(v,w) \mapsto v+w$, and a binary operation called scalar multiplication, $\cdot : \mathbb{C} \times V \to V$, $(\lambda, v) \mapsto \lambda \cdot v$, which obey the following two conditions: $(V,+)$ is a commutative group whose unit element we call 0, and the scalar multiplication obeys

$$\begin{aligned} (\lambda +_{\mathbb{C}} \mu) \cdot v &= \lambda \cdot v + \mu \cdot v, & \lambda \cdot (v+w) &= \lambda \cdot v + \lambda \cdot w, \\ \lambda \cdot (\mu \cdot v) &= (\lambda \cdot_{\mathbb{C}} \mu) \cdot v, & 1_{\mathbb{C}} \cdot v &= v \end{aligned} \tag{2.1}$$

for all $\lambda, \mu \in \mathbb{C}$ and $v, w \in V$, where operations with the subscript $\mathbb{C}$ denote operations on $\mathbb{C}$.

Furthermore, we need to define a vector multiplication ϕ_M on $\mathcal{V}$. We choose to denote $\phi_M : V \times V \to V$ by $(v,w) \mapsto vw$ and demand it to be distributive, as described in the context of a ring above. Furthermore, w.r.t the scalar multiplication, we demand it to obey

$$\lambda \cdot (ab) = (\lambda \cdot a)b = a(\lambda \cdot b). \tag{2.2}$$

Thus, the set V is a ring w.r.t to the vector addition and the vector multiplication. It is called a **complex algebra**.

A complex algebra is therefore nothing more than a complex vector space with an associative and distributive multiplication. Since it is a ring,

- an unit element with respect to the multiplication does not necessarily have to exist.
- an inverse w.r.t the multiplication does not necessarily have to exist.

Those properties will become important in the following sections. Note that some authors, e.g. [Fil05, HS85], demand a complex algebra to contain a unit element.

An example for a complex algebra is the set of 2×2-matrices with complex entries, $Mat(2 \times 2, \mathbb{C})$. Here, a unit element w.r.t the multiplication exists, the unit-matrix, but not every element (i.e., matrix) has an inverse element.

What is a $*$-algebra?

Suppose a complex algebra $\mathcal{A}$ is given. If we define an additional operation on the underlying set A, namely $* : A \to A$, $a \mapsto a^*$, which is *anti-linear*, i.e.

$$(\lambda \cdot a)^* = \bar{\lambda} \cdot a^*$$

($\bar{\lambda}$ denotes a complex conjugation of λ), and where the equations

$$(a^*)^* = a \qquad \text{and} \qquad (ab)^* = b^* a^*$$

hold, $\mathcal{A}$ is called a **$*$-algebra** (or involutive algebra). The element a^* is called *adjoint* to a, and a is called *hermitian* if $a^* = a$.

The above-mentioned example, $Mat(2 \times 2, \mathbb{C})$, is also a $*$-algebra if the involution is defined as complex conjugation and transposition of a matrix.

What is a Banach-$*$-algebra?

Suppose a norm $\|.\|$ is defined on the complex vector space $\mathcal{V}$ used to define a complex algebra $\mathcal{A}$, i.e. a map $\|.\| : \mathcal{V} \to \mathbb{R}$, $v \mapsto \|v\|$, with the properties [HS85]

$$\begin{aligned} \|v\| \geq 0,\ \|v\| = 0 \Leftrightarrow v = 0 \quad & \text{for all } v \in \mathcal{V}, \\ \|\lambda \cdot v\| = |\lambda|\, \|v\| \quad & \text{for all } v \in \mathcal{V}, \lambda \in \mathbb{C}, \\ \|v + w\| \leq \|v\| + \|w\| \quad & \text{for all } v, w \in \mathcal{V}. \end{aligned} \tag{2.3}$$

If furthermore $\|ab\| \leq \|a\|\, \|b\|$ holds, and, if $\mathcal{A}$ possesses a unit element e, then $\|e\| = 1$, we call $\mathcal{A}$ a normed algebra.

If the normed algebra $\mathcal{A}$ is *complete* w.r.t to its norm $\|.\|$, i.e. if every Cauchy sequence has a limit in $\mathcal{A}$, $\mathcal{A}$ is called a Banach algebra or, if an involution is given, a Banach-$*$-algebra.

What is a C^*-algebra?

Given the things said above, there is only one step left to define a C^*-algebra. Namely, a **C^*-algebra** is a Banach-$*$-algebra where the condition

$$\|a^* a\| = \|a\|^2. \tag{2.4}$$

holds.

Example 2.1: Functions on a phase space

In this example, we illustrate the notion of a C^*-algebra by showing that the complex bounded functions on space X form a C^*-algebra.

Let X be a space. Let A be the set of all bounded complex functions on X, i.e.

$$A := \{f \mid f : X \to \mathbb{C},\ \sup_{x\in X}\{f(x)\} < \infty\}.$$

Addition, multiplication and scalar multiplication can be defined in a natural way 'pointwise' by

$$(\alpha \cdot f + g)(x) := \alpha f(x) + g(x),$$
$$(fg)(x) := f(x)g(x)$$

for all $f, g \in \mathcal{A}$ and $\alpha \in \mathbb{C}$. Those definitions can easily be verified to obey Eqs. (2.1), (2.2) and distributivity. Thus, together with those definitions, A is a *complex algebra*, which we denote by $\mathcal{A}$. Note that $\mathcal{A}$ has a unit element, namely the function $f(x) = 1$.

An involution is given on $\mathcal{A}$ by complex conjugation, i.e. by

$$f^*(x) := \bar{f}(x),$$

where $\bar{f}(x)$ is the complex conjugate of $f(x)$. Thus naturally $(\lambda \cdot f)^*(x) = (\bar{\lambda} \cdot f^*)(x)$ and $(f^*)^* = f$. Since furthermore $(fg)^*(x) = \bar{f}(x)\bar{g}(x) = \bar{g}(x)\bar{f}(x) = g^*f^*$, $\mathcal{A}$ is a $*$-algebra.

Let us define the so-called uniform norm (or supremum norm) on A,

$$\|f\| := \sup_{x\in X} |f(x)|,$$

which can easily be proven to obey Eqs. (2.3). It only provides a norm because we work with the set of *bounded* functions on X, which guarantees that $\|f\|$ exists for all $f \in A$. Clearly, $\|1\| = 1$. Since furthermore

$$\|fg\| = \sup_{x\in X} |f(x)g(x)| \leq \sup_{x\in X} |f(x)| \sup_{x\in X} |g(x)| = \|f\|\,\|g\|,$$

$\mathcal{A}$ is a normed $*$-algebra.

To show that $\mathcal{A}$ is complete with respect to $\|.\|$, we take the definition of a Cauchy sequence $(f_n)_{n\in\mathbb{N}}$,

$$\forall\, \varepsilon > 0\ \ \exists\, N \in \mathbb{N}\ \ \forall\, m, n \geq N\ :\ \|f_n - f_m\| < \varepsilon,$$

and use the fact that by definition of the uniform norm this is equivalent to

$$\forall\, \varepsilon > 0 \;\; \exists\, N \in \mathbb{N} \;\; \forall\, m, n \geq N \;:\; |f_n(x) - f_m(x)| < \varepsilon \;\; \forall\, x \in X.$$

The latter statement defines a Cauchy sequence in $\mathbb{C}$, and since we know $\mathbb{C}$ to be complete, we know that for every $x \in X$, the sequence $(f_n(x))_{n\in\mathbb{N}}$ converges to a limit $f(x) \in \mathbb{C}$. This, in turn, guarantees that a $f \in A$ exists with

$$\forall\, \varepsilon > 0 \;\; \exists\, N \in \mathbb{N} \;\; \forall\, n \geq N \;:\; \|f_n - f\| < \varepsilon,$$

i.e. that $(f_n)_{n\in\mathbb{N}}$ converges to f. This shows that every Cauchy sequence in A has a limit in A.

Therefore, $\mathcal{A}$ is a Banach-$*$-algebra.

Last but not least, we need to show that the C^*-condition (2.4) holds:

$$\|f^* f\| = \sup_{x\in X}\{|\bar{f}(x)f(x)|\} = \sup_{x\in X}\{|f(x)|\,|f(x)|\} = \sup_{x\in X}|f(x)| \;\sup_{x\in X}|f(x)| = \|f\|^2.$$

Therefore, the set of bounded complex function on a space X forms a C^***-algebra**.

Representation of an algebra

We conclude this introduction to (C^*-)algebras with some remarks about what a representation of an algebra is. To define it, we need the notion of a morphism between two algebras.

A morphism from a (complex) algebra $\mathcal{A}$ to a (complex) algebra $\mathcal{B}$ is nothing more than map $\varphi : \mathcal{A} \to \mathcal{B}$ which respects the algebraic structure, i.e. $\varphi(\alpha \cdot a + b) = \alpha \cdot \varphi(a) + \varphi(b)$ and $\varphi(ab) = \varphi(a)\varphi(b)$ for all $a, b \in \mathcal{A}$, $\alpha \in \mathbb{C}$ [BEH99].

If a morphism φ is bijective, we call it an *isomorphism*. If $\mathcal{A}$ and $\mathcal{B}$ are $*$-algebras and φ preserves the involution, i.e., $\varphi(a^*) = \varphi(a)^*$, it is called a $*$-*morphism*.

The representation of an algebra $\mathcal{A}$ is a morphism from $\mathcal{A}$ to a set of operators on a suitable space, e.g. a Hilbert space $\mathcal{H}$. If we denote the latter by $\mathrm{Op}(\mathcal{H})$, a representation of $\mathcal{A}$ is thus given by a map

$$\pi : \mathcal{A} \to \mathrm{Op}(\mathcal{H})$$

which has the above-mentioned properties of a morphism. The representation of $\mathcal{A}$ on $\mathrm{Op}(\mathcal{H})$ is called *faithful*, if π is injective.

2.1.2 Physics and C^*-algebras

The appearance of C^*-algebras can be traced back to Gelfand and Neumark. In a 1943 paper [GN43], they defined what is today known as a C^*-algebra [Lan98]. It was known beforehand [vN96] that observables in quantum-theory are given by self-adjoint (hermitian) operators on Hilbert spaces. Thus, when Gelfand and Neumark proved that every C^*-algebra is isomorphic to the norm-closed (i.e., complete) $\star$-algebra of operators on a Hilbert space [GN43, Th. 1], this set the cornerstone for the connection between quantum theory and C^*-algebras [Lan98]. It allowed to understand the observables of quantum-theoretic systems as elements of an abstract algebra which is only represented (c.f. p. 31) by operators on a Hilbert space.

Indeed, the observables of quantum theory are often postulated to form a C^*-algebra in the first place [Fil05]. However, only hermitian elements thereof, obeying $a^* = a$, are considered to correspond to realizable measurements.

The same can be said about observables of classical physical systems. Since they are given by (bounded) functions on a suitable configuration space [Fil05], example 2.1 applies to this case and shows that the set of observables is a C^*-algebra. The real valued functions, which may be considered to be the only "real" observables, are again given by the hermitian elements of this C^*-algebra.

From this point of view, the only difference in the algebraic characterization of observables of both quantum theoretic and classical physical systems is that the former do not commute in general, whereas the latter do.

Summarized, we can say that the observable of both classical physical systems and quantum systems form a C^*-algebra. As mentioned above, this generates the question of whether this is the case for all physical systems. To answer it, observables of NNs are constructed Sec. 2.3, and their properties are analyzed in Sec. 2.4. In Sec. 2.4.1, we evaluate whether the observables of a NN form a C^*-algebra.

States

Before moving on, we make a remark about the notion of 'state' associated with C^*-algebras. The following explanations are taken from [Fil05].

A state is something which determines the results of measurements. Since measurements are represented by observables, i.e. by abstract mathematical objects, a state associates a result with each observable. Therefore, mathematically, a state is a functional which maps every observable of the system to a number. If we denote the observables of a system by $\mathcal{A}$ and chose the number to be complex, a state w is given by

$$w : \mathcal{A} \to \mathbb{C}.$$

A different state consists of a different association of numbers and observables, hence, of a different functional.

Some additional requirements are necessary for a functional to be a state. Namely, the functional is demanded to be positive, normalized and linear. For observables $A, B \in \mathcal{A}$ and $\alpha \in \mathbb{C}$, this means that

$$\begin{aligned} &w(A^*A) \geq 0, \\ &w(1) = 1, \\ &w(\alpha \cdot A + B) = \alpha w(A) + w(B). \end{aligned}$$

The reasons for demanding positivity and normalization can be traced back to quantum theory. There, the number $w(A)$ is not the result of an individual measurement but the expectation-value of the observable A if the system is in the state w. Thus, it is natural to demand that a positive observable has a positive expectation value (positivity) and that the expectation-value of an observable which is represented by the unit-observable is 1 (normalization).

However, the necessity of linearity cannot be motivated like this. Indeed, it was shown by Grete Hermann and later by John S. Bell that linearity does not necessarily have to hold for physical theories. Further information about this can be found in [Fil05].

If the observables of a system form a C^*-algebra, the states of the system (as defined above) can be constructed out of the observables themselves. This is called GNS-construction. An introduction into this topic can be found in [Mat98].

2.2 Motivation 2: Generalized Quantum Theory

We have mentioned above that the question of which structure (i.e., which axioms) can be assumed to hold for arbitrary (physical or non-physical) systems is under current investigation.

One approach along those lines is given by 'Generalized Quantum Theory' (GQT, previously 'Weak Quantum Theory') [ARW02, AFR06, FR11]. Its particular motivation is given by findings in cognitive sciences. Namely, since several years, the results of a number of experiments have successfully been modeled with the mathematical framework of quantum theory [BB12, AR12] (c.f. Sec. 5.1).

The idea behind GQT is to generalize (i.e, weaken) the axiomatic structure of quantum theory in order to allow for a consistent description of phenomena outside the realm of quantum theory. It offers a minimal axiomatic scheme onto which, depending on the phenomenon to be described, additional structure can be added.

GQT is similar to ideas introduced in Sec. 2.1 in that it postulates a set of observables to describe measurements on a system. Most of the axioms of GQT, which will be reviewed in Sec. 2.2.1, constrain the structure of this set of observables. However, those constraints are less

restrictive than a C^*-algebra: E.g., no addition of observables is required to exist. Thus, GQT is a generalization of the concepts of a C^*-algebra.

Compared to conventional classical descriptions of systems and measurements thereon, one of the most important aspects of GQT is that the influences of measurements onto the system under investigation are included into the mathematical description.

Since NNs under experimental inquiry are prime examples of systems subject to state-changing measurements (c.f. Sec. 1.2 and the introduction), they can be used to evaluate whether an approach as suggested by GQT is feasible and whether the axioms of GQT hold in the proposed from.

In the following subsection, we will introduce the axiomatic structure proposed by GQT.

Needless to say that GQT does not assume genuine quantum-physical phenomena to occur outside the realm of quantum physics.

2.2.1 Axioms of GQT

In this section we review the axiomatic structure of GQT. This whole subsection is based on [AFR06, Sec. 2] and, aside from small alterations in notation, all axioms, as well as the quoted comments, are taken from this reference. For slightly alternating view-points and many interesting examples, see e.g. [ARW02] or [FR11].

We assume a *system* Σ to be the object of description. Being non-trivial, this system will have different states z forming a set of states Z. An observable is "a property of the system which can be investigated in a given context" [AFR06, p.35]. Let $\mathscr{A}$ be the set of observables that can sensibly be investigated in this context.

Axiom I To every observable $A \in \mathscr{A}$ belongs a set $specA$, the set of possible outcomes of a "measurement" of A.

Axiom II Observables are (identifiable with) mappings $A : Z \to Z$.

Axiom II expresses the "active, operational character" [AFR06, p.36] of measurements.
Let AB be defined as first observing B, then A.

Axiom III With A and B, AB is an observable as well.

Axiom III implies the associativity of compositions of observables, i.e. $A(BC) = (AB)C$.

Axiom IV There is a unit observable $\mathbb{1}$ such that $\mathbb{1}A = A\mathbb{1} = A \ \forall A \in \mathscr{A}$.

Axiom V There are a zero state o and a zero observable $\mathbb{O}$ such that

$$\begin{aligned}\mathbb{O}(z) &= o \ \ \forall z \in Z,\\ A(o) &= o \ \ \forall A \in \mathscr{A},\\ A\mathbb{O} &= \mathbb{O}A = \mathbb{O} \ \ \forall A \in \mathscr{A}.\end{aligned}$$

Define a proposition P to be an observable ($\neq \mathbb{O}, \mathbb{1}$) "whose outcome is either true or false" [AFR06, p.36], i.e.

$$specP = \{true, false\}.$$

Let P^c denote the negation of P. I.e., $P^c(z)$ is an observable which has exactly the opposite truth-values of $P(z)$. Two propositions P_1 and P_2 are called compatible, if $P_1P_2 = P_2P_1$. For compatible propositions, the conjunction $P_1 \wedge P_2$ and the adjunction $P_1 \vee P_2$ are defined as

$$\begin{aligned}P_1 \wedge P_2 &= P_1P_2 = P_2 \wedge P_1,\\ P_1 \vee P_2 &= (P_1^c P_2^c)^c = P_2 \vee P_1.\end{aligned}$$

Axiom VIa Propositions satisfy the following conditions:

$$\begin{aligned}P^2 &= P,\\ (P^c)^c = P,\ \mathbb{1}^c &= \mathbb{O},\\ PP^c = P^cP &= 0.\end{aligned}$$

Axiom VIb If $P(z) \neq o$, then $P(z)$ is a state in which P is true with certainty.

For $\alpha \in specA$ let A_α denote the proposition that the outcome of the measurement of A is α.

Axiom VIc Let

$$\begin{aligned}A_\alpha A_\beta = A_\beta A_\alpha = \mathbb{O} \ \text{ for } \alpha \neq \beta,\\ AA_\alpha = A_\alpha A,\\ \bigvee_{\alpha \in specA} A_\alpha = \mathbb{1}.\end{aligned}$$

Note two things:

1. Neither on the set of observables $\mathscr{A}$, nor on the set of states Z, addition is defined. (Of course, it can be amended if necessary for a specific system.)

2. It is essential that commutativity of observables, i.e. $AB = BA$ for $A, B \in \mathscr{A}$, is *not* assumed. In general, there will be commutative and non-commutative observables.

The following key features of ordinary quantum theory are not included in or cannot be derived from the axioms of GQT: the Schrödinger equation, the Born rule, a Hilbert space structure, the von Neumann algebra, the Heisenberg uncertainty principle, the Bell inequalities.

An evaluation in how far NNs obey the axioms of GQT can be found in Sec. 2.4.2.

2.3 Observables on NNs

In this section, we construct observables of NNs. To do so, we take the definitions of Sec. 1.1 and of Sec. 1.2 and construct an algebraic representation of the process of measurement. I.e., every measurement will be represented by a matrix which acts on a vector-space, where the vectors of this space represent the states of the NN under investigation. This allows to study the mathematical structure of the set of observables of NNs in Sec. 2.4, and particularly to evaluate the questions posed by the last sections.

The matrices which occur in the following will take a familiar form: they are Markov matrices (also called stochastic matrices). This is not surprising since we have specified measurements to be evaluated only when the NN has settled into an attractor in Sec. 1.2.

We will use the following concepts of previous sections: the closed set of attractors, the attractor correspondence and the generalized output-function. In order to define the closed set of attractors and the attractor-correspondence, we have assumed that a set of inputs, denoted by $\mathcal{I}$, is specified. Thus, the following considerations also make this assumption: The set of observables, which is constructed in the following, depends on the set of inputs $\mathcal{I}$, which is specified.[1]

In Sec. 1.1.5 (p. 17), we have defined the attractor-correspondence $f(i, z)$ to consist of all attractors which a NN may reach if input i is presented to the NN while it occupies attractor z. Furthermore, we have derived probabilities associated with each possibility (Eq. (1.26)). If $\mathcal{Z}$ consists of N attractors, i.e. $\mathcal{Z} = \{z_1, ..., z_N\}$, this allows us to represent the NN's behavior in the following way:

$$f(i, z_k) = \begin{cases} z_1 & \text{with } p_i(z_1|z_k) \\ ... & \\ z_N & \text{with } p_i(z_N|z_k). \end{cases} \tag{2.5}$$

This equation describes how the NN behaves in terms of attractors once a measurement is performed: It represents the process of making a measurement abstractly.

For a fixed input i, the attractor-correspondence takes an attractor of $\mathcal{Z}$ and associates it with

[1] This is so in a twofold way. First, since every input corresponds to a different measurement which can be performed on the NN, and since every measurement is represented by an observable, the set $\mathcal{I}$ specifies which observables appear. Second, observables are operators (or matrices) on a vector space which is spanned by the attractors of a closed set of attractors. If another input-set is specified, the closed set of attractors is different (c.f. Sec. 1.1.4), and thus, the vector space may change.

other attractors of $\mathcal{Z}$. Thus, it is similar to an operator on $\mathcal{Z}$, which motivates the definition

$$A_i\, z := f(i, z). \tag{2.6}$$

Thus, every input i corresponds to one A_i. Strictly speaking, the term 'operator' is not justified in this case because A_i may associate more than one attractor of $\mathcal{Z}$ to some attractors of $\mathcal{Z}$. To make this explicit, we can rewrite Eq. (2.5) as

$$A_i\, z_k = \begin{cases} z_1 & \text{with } p_i(z_1|z_k) \\ ... & \\ z_N & \text{with } p_i(z_N|z_k). \end{cases} \tag{2.7}$$

For later use, we define the composition of A_i and A_j as

$$A_i A_j\, z = f(i, f(j, z)). \tag{2.8}$$

It is clear that Eq. (2.7) is not a suitable mathematical object to evaluate, e.g., whether the $A_i A_j = A_j A_i$. Therefore, we seek a suitable representation of the Eq. (2.7) in the following, which allows to study the properties of measurements. Those representations, as mentioned above, are then observables of the NN: They represent measurements in a suitable, abstract way.

In order to construct this representation, we will introduce a vector space with a scalar product. The idea behind this is that if we associate each attractor of $\mathcal{Z}$ with a basis-vector of this vector-space, the right-hand-side of Eq. (2.7) can be represented by a vector of this vector-space: by a linear combination of basis-vectors with coefficients $p_i(z_1|z_k)$, etc.

This has important consequences: Since A_i is well-defined for all $z \in \mathcal{Z}$, it acts in a natural way on vectors of the vector-space: It takes one of those vectors and maps it to one vector of the vector space. Since it only associates one vector of this vector-space to every other vector, it is indeed an operator (or matrix) on this vector-space, and the conventional methods to analyze matrices can be applied. I.e., with this construction, we can investigate the properties of measurements by using the well-known methods linear algebra, which is the essential motivation for introducing observables in the first place.

For $\mathcal{Z} = \{z_1, ..., z_N\}$, choose an **$N$-dimensional $\mathbb{R}$-vector-space**, on which a positive definite, non-degenerate and symmetric **scalar product**[2] $\langle \cdot, \cdot \rangle$ is defined. (Technically this is called a $\mathbb{R}$-Hilbert space [HS85]. However, some authors reserve this term for infinite-dimensional vector spaces.)

We will denote this vector-space by V.

As explained above, we associate an (orthonormal) basis vector v_k of V with each attractor $z_k \in \mathcal{Z}$.[3] Therefore, we can interpret all vectors v of the form

$$v = \alpha_1 v_1 + \alpha_2 v_2 + ... + \alpha_N v_N \tag{2.9}$$

[2] A scalar product is sometimes also called dot-product or inner product. It takes two vectors and maps them to a real number.

[3] In the following, let v_k always refer to this particular basis-vector of V, which represents attractor $z_k \in \mathcal{Z}$.

with

$$\sum_{i=1}^{N} \alpha_i = 1 \qquad \text{and} \qquad \alpha_i \in [0,1] \subset \mathbb{Q} \tag{2.10}$$

as the NN being in attractor v_i with probability α_i $(i = 1, ..., N)$. The constraints (2.10) appear because the probabilities on the right hand side of (2.7) add up to one (c.f. Eq. (1.27)), and because they can only be rational numbers by construction. Both of those properties will be of importance later.

In light of this, we can understand Eq. (2.9) to represent a state of the NN. Therefore, we call vectors $v \in V$ of the form (2.9) which obey Eqs. (2.10) **NN-state-vectors**. We define the set of all NN-state-vectors in V as $\mathcal{S}$.

Thus, a NN-state-vector represents a probability-distribution over the set of attractors $\mathcal{Z}$. It specifies which attractor a NN occupies with which probability.

Those things having been said, we can write Eq. (2.7) as

$$A_i \, v_k = p_i(z_1|z_k) \, v_1 + p_i(z_2|z_k) \, v_2 + ... + p_i(z_N|z_k) \, v_N. \tag{2.11}$$

Since this is possible for every basis-vector v_k ($f(i, z_k)$ specifies the behavior of the NN for every attractor $z_k \in \mathcal{Z}$), Eq. 2.11 defines a matrix. This matrix represents all information contained in $f(i, z)$ for a given input as well as the epistemic probabilities $p_i(z_l|z_k)$ associated with this input. Thus, this matrix is the **observable** of the NN associated with input i. It is given by[4]

$$A_i \doteq \begin{pmatrix} p_i(z_1|z_1) & ... & p_i(z_1|z_N) \\ p_i(z_2|z_1) & ... & p_i(z_2|z_N) \\ ... & ... & ... \\ p_i(z_N|z_1) & ... & p_i(z_N|z_N) \end{pmatrix}. \tag{2.12}$$

Once observables are given, we can comfortably analyze their properties. Consider, e.g., the question of whether a given observable has an inverse. Since the composition of observables is naturally given by matrix multiplication (c.f. Eq. (2.8)), we can determine whether an observable has an inverse by checking if its matrix is invertible. However the existence of an inverse of the matrix representing an observable does not suffice to guarantee the existence of an inverse observable. Rather, we have to check whether the inverse itself is an observable of this NN (or of any NN at all).

In order to do this, we specify which properties necessarily hold for every observable of a NN. This yields the set $\mathcal{N}$ which is defined as follows.

Let $\mathcal{N}$ be the set of $N \times N$-matrices which are interpretable as observables on a NN, i.e.

[4] This is so since

$$A_i v_k \doteq \begin{pmatrix} p_i(z_1|z_1) & ... & p_i(z_1|z_N) \\ p_i(z_2|z_1) & ... & p_i(z_2|z_N) \\ ... & ... & ... \\ p_i(z_N|z_1) & ... & p_i(z_N|z_N) \end{pmatrix} \begin{pmatrix} ... \\ 1 \\ ... \\ ... \end{pmatrix} = \begin{pmatrix} p_i(z_1|z_k) \\ p_i(z_2|z_k) \\ ... \\ p_i(z_N|z_k) \end{pmatrix} \doteq p_i(z_1|z_k) \, v_1 + ... + p_i(z_N|z_k) \, v_N,$$

where all entries of the second vector are 0 except the k^{th}, which is 1.

- whose columns add up to 1 (c.f. Eq. (1.27)),
- whose matrix-elements are elements of $[0, 1] \in \mathbb{Q}$.

Thus the set $\mathcal{N}$ essentially (i.e., asides from the appearance of $\mathbb{Q}$) consists of Markov (or stochastic) matrices. However, the appearance $\mathbb{Q}$ rather than $\mathbb{R}$ at this point is essential. The reason for this is that the probabilities as defined in Eq. (1.26) can only be rational, not irrational numbers.

$\mathcal{N}$ comprises all matrices which can be observables of some NN whose closed set of attractors consists of N attractors ($N > 0$ can be arbitrary but finite). In order to work with those observables, which represent measurements of a particular given NN, we define $\mathscr{I}$ to be the set of $N \times N$-matrices which are observables of a *particular* NN. (Thus $\mathscr{I} \subset \mathcal{N}$.)

Only if the inverse of the matrix representing an observable is part of $\mathscr{I}$, the observable itself has an inverse. And only if the inverse matrix is part of $\mathcal{N}$, there could in principle be a NN for which this observable has an inverse.

In a similar manner, we can check other properties of the observable-algebra, e.g. if observables commute or if an observable corresponding to unity exists.

Characterization of measurements

Based on the notions of NN-state-vectors and observables, we can specify what exactly characterizes a measurement in the representation which has been introduced above.

Suppose the NN is in a NN-state-vector v, and suppose a measurement belonging to input i is performed. The question arises of which output of the NN to expect with which probability.

The answer follows directly from the construction of v and A_i above, c.f. Eqs. (2.9) and (2.12). Namely, since $\langle v_k, v\rangle = \alpha_k$ for v as specified in Eq. (2.9), the probability of the NN settling in attractor z_k if input i is presented while it occupies NN-state-vector V is

$$\langle v_k, A_i v\rangle, \tag{2.13}$$

where $\langle \cdot, \cdot \rangle$ represents the scalar-product on V, as introduced above.

Note that the absence of a non-linearity in Eq. (2.13) implies that there are no interference-effects between different states.

Since NN-state-vectors v represent probability-distributions over the closed set of attractors $\mathcal{Z}$, once the output of a measurement associated with a given input is known, the NN-state vector can be updated. The updated state-vector v' can only be given by a linear combination of those basis-vectors which correspond to an attractor whose output is equal to the one the NN shows. Therefore, v' is given by the projection of v onto the subspace which is spanned by the just-mentioned basis-vectors. We will express this mathematically in the following by defining a suitable projector.

A projector associated with a given output o of the NN can be defined in the following way.

Let $o \in \mathcal{G}_{\mathcal{Z}}$ be an output of the NN. We denote by

$$P_o \tag{2.14}$$

the projector onto the subspace spanned by exactly those basis-vectors whose corresponding attractors have the output o. I.e.,

$$G(o) := \{z_k \in \mathcal{Z} \mid g_{\mathcal{Z}}(z_k) = o\}, \tag{2.15}$$

$$P_o v := \sum_{z_k \in G(o)} \langle v_k, v \rangle v_k, \tag{2.16}$$

where v_k is the basis-vector of V which represents attractor z_k.

Thus the gain in knowledge when reading off the output of a NN corresponds to a change

$$v \longrightarrow \frac{P_o v}{N(P_o v)},$$

where $N(P_o v)$ denotes a normalization of $P_o v$, which ensures that the resulting vector is again a NN-state-vector. I.e., for $v = \sum_{i=1}^{N} \alpha_i v_i$,

$$N(v) := \sum_{i=1}^{N} \alpha_i.$$

Thus, a measurement on a NN occupying the NN-state-vector v is mathematically characterized by the following two points:

a) The result of the measurement can be the output of any attractor $z_k \in \mathcal{Z}$ for which

$$\langle v_k, A_i v \rangle \neq 0,$$

where A_i is the **observable** which represents the measurement algebraically. $\langle v_k, A_i v \rangle$ is the probability to reach v_k from v if input i is presented.

b) Once the measurement has been performed (i.e. the input i has been shown and the output o 'observed'), this corresponds to a change in the NN-state-vector by

$$v \longrightarrow \frac{P_o A_i v}{N(P_o A_i v)}. \tag{2.17}$$

Before a measurement, only probabilistic predictions are possible and the act of observation corresponds to a projection onto the subspace of states with the given measurement-outcome.

We conclude this section by defining a natural notion of commutativity.

Commutativity The natural notion of commutativity suggested by Eq. (2.8) is

$$A_i, A_j \text{ commute } \Leftrightarrow A_i A_j = A_j A_i, \tag{2.18}$$

Hence, if two observables commute, the state-vector of the NN after presenting the inputs i and j is independent of the order of presentation.

Summary 4: Algebraic Treatment of NNs

In Sec. 2.3, we have constructed observables of a NN.

A suitable representation of measurements can be constructed in a N-dimensional $\mathbb{R}$-Hilbert space V (p. 37), where $N = |\mathcal{Z}|$. Basis-vectors v_k of V represent attractors of the closed set of attractors $\mathcal{Z}$, and some linear combinations thereof, called NN-state-vectors (c.f. Eq. (2.9)), represent probability-distributions over $\mathcal{Z}$.

Thus, observables of the NN are given by matrices, such as (2.12).

The probability to find the NN in $z_k \in \mathcal{Z}$ after a measurement associated with input i has been performed on a NN described by NN-state-vector $v \in V$ is given by

$$\langle v_k, A_i v \rangle,$$

where v_k is the basis vector of V which represents z_k. If the output, which the NN displays subsequently, is o, the measurement corresponds to a change in the state-vector of the NN from v to

$$\frac{P_o A_i v}{N(P_o A_i v)},$$

where P_o is defined in Eq. (2.16) and $N(P_o A_i v)$ is a normalization-factor (p. 40).

2.4 General properties of the set of observables of NNs

As elaborated in the last sections, several questions concerning properties of the observables of NNs are to be answered in this thesis, including: Do the observables of a NN respect the structure of a C^*-algebra? When do observables commute? Is a treatment along the lines of GQT feasible? Do the axioms of GQT hold?

With the definitions of Ch. 1 and of Sec. 2.3, the explicit set of observables of a *given* NN, and thus its properties can be calculated. However, the scope of this thesis includes the question of which properties to expect in general. This section offers some results concerning the latter question.

In the following, we will present some lemmata which specify general properties of the set of observables of NNs. We will relate those results to C^*-algebras in Sec. 2.4.1 and to GQT in Sec. 2.4.2.

Lemma 2.1. *In general, the observables of a NN do not commute.*

To prove lemma 2.1 as well as the following lemmata, a single example of a NN whose observables do not commute would suffice. However, proving the non-commutativity for a more general class of NNs will provide a more satisfactory result. Therefore, we will use the class of perfect learners, as defined in Sec. 1.3.3, to prove the following lemmata.

Proof Suppose a perfect learner is given as defined in Sec. 1.3.3. Assume for simplicity that there exists at least one attractor z_m in $\mathcal{Z}$ which has a unique output. Let i be an input which leads to z_m, which has to exist due to the minimal closure of $\mathcal{Z}$ (c.f. p. 15). Thus:

$$\exists\, z_l \,:\, z_m \in f(i, z_l).$$

Eq. (1.32) gives

$$g_{\mathcal{Z}}(f(i, z_l)) = g_{\mathcal{Z}}(f(i, z_k)) \qquad \forall\, z_k \in \mathcal{Z} \quad \forall\, i \in \mathcal{I}.$$

Together with the assumption of z_m possessing a unique output, this implies that

$$\{z_m\} = f(i, z_l) = f(i, z_k) \qquad \forall\, z_k \in \mathcal{Z} \quad \forall\, i \in \mathcal{I}.$$

By construction of $p_i(z_m|z_k)$ (p. 18), this yields

$$\begin{aligned} p_i(z_m|z_k) &= 1 \qquad \forall\, z_k \in \mathcal{Z},\, \forall\, i \in \mathcal{I} \quad \text{and} \\ p_i(z_n|z_k) &= 0 \qquad \forall\, z_k \in \mathcal{Z},\, z_n \in \mathcal{Z}\backslash\{z_m\},\, \forall\, i \in \mathcal{I}. \end{aligned}$$

With Eq. (2.12), this gives

$$A_i \doteq \begin{pmatrix} 0 & 0 & ... & 0 & 0 \\ ... & ... & ... & ... & ... \\ 1 & 1 & ... & 1 & 1 \\ ... & ... & ... & ... & ... \\ 0 & 0 & ... & 0 & 0 \end{pmatrix}, \tag{2.19}$$

where all matrix-elements are 0 except for the m^{th} row which consists of 1s. For every NN which has more than one possible output ($|\mathcal{G}_{\mathcal{Z}}| > 1$), there are inputs $j \in \mathcal{I}$, for which the perfect learner property gives

$$p_j(z_m|z_k) = 0 \qquad \forall\, z_k \in \mathcal{Z},$$

which implies that the m^{th} row of the representation of observables associated with those inputs consists of 0s only. This suffices to guarantee that their representing matrices do not commute with A_i, and hence that A_i and A_j do not commute.

Q.e.d.

Lemma 2.2. *In general, the observables of a NN are not invertible.*

Proof Consider again observable A_i of the perfect learner treated in proof of lemma 2.1 as given in Eq. (2.19). This matrix is not invertible.[5]

Q.e.d.

Lemma 2.3. *In general, there is no unit element among the observables of a NN.*

Proof We consider again the class of perfect learners. If a unit-observable A_k would exist, it would have to obey

$$p_k(z_i|z_i) = 1 \qquad \forall\, z_i \in \mathcal{Z}. \tag{2.20}$$

This is in conflict with property (1.32), since the output of such an observable would clearly not be independent of the state of the system.

Q.e.d.

2.4.1 A C^*-algebra?

In this section, we evaluate whether the set of observables of NNs, defined as $\mathcal{N}$ on p. 38, is a C^*-algebra.

We have introduced the notion of a C^*-algebra in Sec. 2.1. Among the details of the definition, the most important requirement for a set to form a C^*-algebra (indeed an algebra at all) is the existence of three binary relations: addition, multiplication and scalar multiplication (c.f. Sec. 2.1.1).

If we look at the set $\mathcal{N}$ under investigation, it is clear that such operations exist in a natural way. Namely, since $\mathcal{N}$ contains matrices, they are given by matrix-multiplication, matrix-addition and scalar multiplication with a matrix.

This allows us to give a first answer to the question raised in this section:

Lemma 2.4. *The observables of a NN as given by $\mathcal{N}$ do not form a C^*-algebra in a natural way.*

[5] An easy way to see this is to consider the formula

$$\det A = \sum_{i_1,\ldots,i_N=1}^{N} \varepsilon_{i_1 \ldots i_N}\, a_{1,i_1} \cdots a_{N,i_N},$$

where A is a $N \times N$-matrix, $\varepsilon_{i_1 \ldots i_N}$ is the N-dimensional totally antisymmetric Levi-Civita-symbol and $a_{j,k}$ are the matrix-elements of A. It follows directly that $\det A_i = 0$ for A_i given by Eq. (2.19), and thus that A_i is not invertible [RHB06].

Proof For the following reasons, the natural binary relations on $\mathcal{N}$ do not define a C^*-algebra:

- The natural definition of addition on the set of observables $\mathcal{N}$ is the matrix addition. With respect to this definition, the set of observables is not closed, i.e. the addition of two matrices of the set $\mathcal{N}$ can yield a matrix which is not in $\mathcal{N}$. This is obvious, e.g., when adding a matrix to itself.
- The 0-matrix, whose entries are all equal to 0, is not part of the observables of $\mathcal{N}$ as defined on p. 38. Thus with respect to a natural addition on $\mathcal{N}$, no 0 exists.
- Similarly, the set $\mathcal{N}$ is not closed with respect to natural scalar multiplication. If a matrix of $\mathcal{N}$ is multiplied by a number $\lambda \neq 1$, the resulting matrix is not in $\mathcal{N}$ any more.

Q.e.d.

However, one might wonder if there is a different definition of binary relations on $\mathcal{N}$ which *does* make $\mathcal{N}$ a C^*-algebra. After all, the essence of the introduction of an algebra associated with observables is that it captures properties of the latter *independent* of any representation (c.f. Sec. 2.1.2).

E.g., one could define an addition on $\mathcal{N}$ to consist of matrix addition combined with a suitable normalization which does make the resulting matrix an element of $\mathcal{N}$. But one could also define an even more unnatural notion of addition or (scalar) multiplication on $\mathcal{N}$. Of course, for every choice, one would have to show that the definitions of Sec. 2.1.1 do hold, e.g., that the chosen binary relations are distributive.

This leads us to the question of whether there is a general argument which prevents the set $\mathcal{N}$ to form a C^*-algebra regardless of the specific choice of the definition of binary relations on $\mathcal{N}$. Any such argument would have to show that some part of the definition of a C^*-algebra cannot be fulfilled by $\mathcal{N}$ regardless of the definitions of binary relations.

In the following, we will present such an argument. We will show that the set $\mathcal{N}$ cannot be complete, i.e., that there are Cauchy sequences in $\mathcal{N}$ which do not have a limit in $\mathcal{N}$ (c.f. p. 29). This argument rests essentially on the appearance of $\mathbb{Q}$ in the definition of $\mathcal{N}$ (c.f. p. 38) since it is $\mathbb{Q}$ itself which is not complete.

We emphasize again that this does *not* mean that only due to appearance of $\mathbb{Q}$, $\mathcal{N}$ does not form a C^*-algebra. Indeed, it is unlikely that any definition of binary operations can be found at all which satisfies all other definitions of a C^*-algebra. The completeness-requirement is just a convenient property which allows a prove of the above point.

The completeness (in the above sense) of mathematical structures associated with physical systems is a delicate matter. On one hand, it allows mathematical statements to be constructed which are otherwise not possible, but on the other hand completeness cannot be verified in experiments. The reason for this is that the notions of Cauchy sequences and of convergence refer to infinite sequences, which cannot be realized in experiments.

However, since this section merely aims to evaluate in how far the set $\mathcal{N}$ is a C^*-algebra, and since the completeness contained in the definition of the latter, we can legitimate state and prove the following theorem.

Theorem 1 C^***-property**
The observables of a NN, $\mathcal{N}$, do not form a C^*-algebra.

Proof As noted above, an essential point in the definition of the set $\mathcal{N}$ (p. 37) was that $\mathcal{N}$ only contains matrices whose entries are rational numbers. In the following, we will use this property to show that $\mathcal{N}$ cannot form a complete space. In accordance with Sec. 2.3, let N denote dimension of the Hilbert space on which the observables of a NN are represented.

Since the rational numbers do not form a complete metric space, there exists a Cauchy sequence $(b_n)_{n\in\mathbb{N}} \subset [0,1] \subset \mathbb{Q}$ which does not have a limit in $\mathbb{Q}$, i.e.

$$b_n \not\to b \in \mathbb{Q}.$$

We use this sequence of numbers to construct a sequence of matrices $(A_n)_{n\in\mathbb{N}} \in \mathcal{N}$. Let a^n_{ij} denote the matrix-elements of A_n and define

$$\begin{aligned} a^n_{11} &:= b_n && \forall\, n \in \mathbb{N}, \\ a^n_{21} &:= 1 - b_n && \forall\, n \in \mathbb{N}, \\ a^n_{ii} &:= 1 && \forall i \in \{2, ..., N\}\, \forall\, n \in \mathbb{N}, \end{aligned}$$

and $a_{ij} := 0$ otherwise.

The Hilbert space V, on which matrices of $\mathcal{N}$ act (p. 37), is finite dimensional, which is a consequence of the finiteness of $\mathcal{Z}$ (c.f. p. 15). Thus the space of *linear* transformations of V (the space of matrices on V) is also finite dimensional.

This has important consequences for our purposes. Namely, concerning convergence, all norms of a finite dimensional vector space are equivalent [BSMM05, Sec. 12.3.1.2]. Thus, we can choose an arbitrary norm to show whether a subset of matrices on V is complete or not, e.g. $\|A\|_{\max} := \max\{|a_{ij}|\}$.

$(A_n)_{n\in\mathbb{N}}$ is a Cauchy sequence since the definition of a Cauchy sequence [HS85],

$$\forall\, \varepsilon > 0 \;\; \exists\, N \in \mathbb{N} \;\; \forall\, m, n \geq N \;:\; \|A_n - A_m\|_{\max} < \varepsilon,$$

is, given the above choice of a norm, equivalent to

$$\forall\, \varepsilon > 0 \;\; \exists\, N \in \mathbb{N} \;\; \forall\, m, n \geq N \;:\; |b_n - b_m| < \varepsilon,$$

which holds by construction of $(b_n)_{n\in\mathbb{N}}$. However, since $b_n \not\to b \in \mathbb{Q}$,

$$A_n \not\to A \in \mathcal{N}.$$

Thus we have shown that there exists a Cauchy sequence in $\mathcal{N}$ which does not converge in $\mathcal{N}$, and thus that $\mathcal{N}$ is not complete.

Therefore, $\mathcal{N}$ *cannot be* a C^*-algebra regardless of any definitions of binary operations on $\mathcal{N}$.
Q.e.d.

Before concluding this section, we address the question of whether the set $\mathcal{N}$ might be a subset of a C^*-algebra in a natural way, similar to the quantum-theoretic observables, which are hermitian elements of a C^*-algebra. This question motivated by the fact that the rational numbers $\mathbb{Q}$ lie dense in $\mathbb{R}$, where the latter is complete in the sense relevant in this section.

An answer to this question depends strongly on what one understands as 'natural'. At this point, we only refer to the proof of lemma 2.4, which shows that with respect to natural binary operations, an embedding would have to be unnatural.

2.4.2 NNs and GQT

This section relates the results, which have been obtained above, to GQT, which has been introduced in Sec. 2.2.

Sec. 2.3 shows that a treatment of NN along the lines of GQT, as introduced in Sec. 2.2, is in principle feasible. However, not all axioms of GQT hold. In the following we will associate the latter with the notions developed in Sec. 2.3 and 2.4.

To evaluate the axioms of GQT (Sec. 2.2.1), we define the system Σ under investigation to be a NN. In accordance with Sec. 2.3, we define the *states* of Σ to be the NN-state-vectors as defined on p. 37. Thus $Z = \mathcal{S}$. Observables correspond to measurements, as defined in Sec. 1.2, and the set of observables of a given NN is given by $\mathcal{J}$, as defined on p. 38.

Thus, we can associate a set of possible outcomes of a measurement with every observable $A_i \in \mathcal{J}$. For a particular observable A_i it is given by the output of those attractors which can be reached from an arbitrary attractor subsequently to a presentation of input i to the NN. I.e., it is given by

$$\{g_{\mathcal{Z}}(z') \mid \exists\, z \in \mathcal{Z} \,:\, z' \in f(i,z)\},$$

which implies that axiom I is satisfied. Since observables correspond to mappings $\mathcal{S} \to \mathcal{S}$, the same is true for axiom II.

Concerning the Hilbert space representation of observables, if A_i and A_j are observables, also $A_j A_i$ is an observable. It is realized by showing input i, then input j and registering the output afterwards. Thus, axiom III is satisfied.

Lemma 2.3 shows that there is no unit element among the observables of a NN in general. This implies that axiom IV is not satisfied.

A zero state o, as described in axiom V, can only be given by the 0-vector of V. Since this vector is not a NN-state-vector, there is no such state in $\mathcal{S}$. The same is true for a zero observable $\mathbb{O}$. Thus, axiom V is not satisfied.

Since an observable with only two possible outputs is generally not given on a NN, there is no observable - in the above sense - corresponding to a proposition, as defined in axioms VI. Therefore, in general, axiom VI is not satisfied.

Summarized, axioms I to III of GQT are satisfied by NNs, whereas axioms IV, V and VI are not satisfied.

Summary 5: NNs, GQT and C^*-algebras

After having introduced C^*-algebras in Sec. 2.1 and the axiomatic scheme of Generalized Quantum Theory (GQT) in Sec. 2.2, we have studied general properties of observables of NNs in Sec. 2.4.

We have proven that the observables of NNs do not commute in general (lemma 2.1, p. 42) and that an inverse of an observable does not exist in general (lemma 2.2, p. 43). Furthermore, there is no unit element among the observables of a NN in general (lemma 2.3, p. 43).

With respect to GQT, this chapter has a twofold implication. First, Sec. 2.3 shows that NNs can be treated as proposed by GQT. Second, the lemmata of Sec. 2.4 show that not all of the axioms postulated by GQT hold on NNs. Particularly, axioms III to VI are violated.

In Sec. 2.4.1, we have shown that the observables of NNs, as given by $\mathcal{N}$, do not form a C^*-algebra in a natural way. (p. 43)

Additionally, we have offered a theorem which shows that regardless of any definitions of binary operations on $\mathcal{N}$, the latter cannot form a C^*-algebra (p. 45). Therefore, Sec. 2.4.1 suggests that the question of whether the observables of any physical system form a C^*-algebra should be negated.

Chapter 3

Identification of attractors

In this chapter, we develop mathematical structures which relate the state of a NN (c.f. below) and series of consecutive measurements. This is of conceptual interest for the following reasons.

Even though a state influences the results of measurements by definition, it is not clear for an arbitrary system how a sensible notion of 'state' can be reconstructed from measurements. E.g., in quantum theory, the question of how a ray in a Hilbert space can be identified is non trivial: any single measurement will only yield information about which eigenvector of an observable the system occupies subsequent to the measurement.

Furthermore, for many systems it is non-trivial to prepare the system in arbitrary states. Here, NNs serve as a good example: It is not clear a priori how a NN can be prepared, e.g., in an arbitrary one of its attractors through measurements.

This motivates an inquiry into the structural relation between states of a NN (to be defined below) and series of consecutive measurements, i.e., histories (c.f. Sec. 1.2.3, p. 22).

We have mentioned before (Sec. 1.2.1) that this chapter focuses on epistemic states, and that a sensible notion of epistemic states is given by the attractors of a NN. However, this clam is only justified in anticipation of the results of this chapter (c.f. Sec. 1.2.1), a priori any of the following notions of 'state', which have appeared throughout this thesis, could be called an epistemic state.

First, we have encountered the *activation* of a NN (p. 8), which specifies the states of all non-input neurons of a NN. Whereas the activation of a NN does determine the dynamics of the system to a certain extend (not completely since the update-rule might depend on previous activations as well), it cannot be identified by measurements.[1] Therefore, it does not constitute a notion of epistemic states.

A second notion of state was introduced in connection with C^*-algebras in Sec. 2.1.2. There, a state has been defined as a linear, positive and normalized functional on the set of observables.

[1] As mentioned before, the reason of why different activations can in general not be discriminated by measurements is twofold: a) Measurements change the state of the NN and b) the output of a NN allows only limited information about the states of the neurons of the NN to be obtained.

However, since the observables of a NN do not form a C^*-algebra, and particularly since a natural notion of addition does not exist on the set of observables $\mathcal{N}$, this notion of state does not represent epistemic states either.

A third notion of state has appeared in connection with GQT (Sec. 2.2). Here, observables are postulated to be mappings from a suitable state-set into itself. We have already shown in Ch. 2 that a treatment as proposed by GQT is indeed feasible, and have found that the set of states is given by probability distributions over a closed set of attractors.

This chapter focuses on the question of how the attractor, which a NN occupies can be identified through series of measurements. Since explicit methods will be constructed, which allow to do so, we can, in anticipation of the results, identify the set of epistemic states of a NN with the closed set of attractors $\mathcal{Z}$. Therefore, an **epistemic state of a NN** is one attractor of $\mathcal{Z}$.

In a first part of this chapter, we assume some information about the NN under investigation to be available, namely the attractor-correspondence of a closed set of attractors $\mathcal{Z}$ (p. 16) and the generalized output-function (p. 16). This implies that a set of inputs $\mathcal{I}$ has to be fixed.

Whereas the question of which attractors of $\mathcal{Z}$ a NN under investigation can occupy after one single measurement is trivial (namely all those attractors $z \in \mathcal{Z}$ whose associated output $(g_{\mathcal{Z}}(z))$ equals the one which the NN displays), the question of how to further constrain the possibly occupied attractors by consecutive measurements is non-trivial. Put differently, the question of how to pick out those attractors of a NN which can be occupied after a series of measurements has been performed is non-trivial.

To answer this question, we introduce a mathematical representation of $\mathcal{Z}$, of the attractor-correspondence and of the generalized output-function in Sec. 3.1. It is given by a graph whose nodes represent elements of $\mathcal{Z}$ (and are colored by the generalized output-function), and whose edges represent the attractor-correspondence. Given this representation, which will be called mapping graph, a relation between the attractors which a NN can occupy and series of measurements can easily be constructed (Sec. 3.2). Particularly, the attractor which a NN has occupied before or after any one of the measurements contained in a history can be specified, and questions of how to prepare a NN in an arbitrary attractor can be answered.

However, it is unlikely that information about the closed set of attractors, the attractor-correspondence and the generalized output function (i.e., about the mapping graph of a NN) is available in an experimental investigation of NNs. Therefore, Sec. 3.3 offers an algorithmic procedure which allows to obtain the mapping graph of a NN (or a subgraph thereof) only through evaluation of the results of measurements on this NN. The drawback of this procedure is that it assumes an ensemble of NNs to be available. However, there are possibilities to avoid this limitation.

Summarized, this chapter offers mathematical possibilities to identify the attractor, which a NN occupies at a given time by performing series of consecutive measurements. The concepts, which are developed, might ultimately be applied to real experiments, c.f. Ch. 4. However, further

research is necessary in this direction.

3.1 Mapping graph

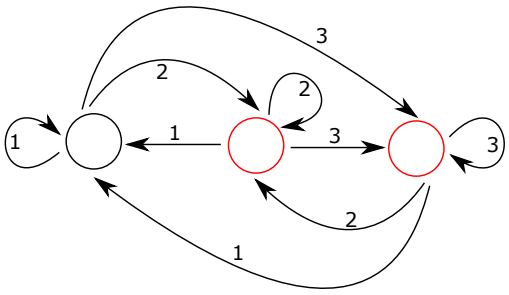
In the following, we introduce a concept which underlies the considerations of chapter 3. It represents the behavior of a NN on the level of attractors. Since it is a representation of the attractor-correspondence $f(i, z)$ defined in Eq. (1.23), it will be called 'mapping graph'. However, in contrast to the representation constructed in Sec. 2.3, it does not represent the probabilities associated with each transition from one attractor to the next. Formally, the mapping graph is a directed, vertex- and edge-colored multigraph.[2]

A **mapping-graph** underlying a given NN is defined by the following conditions:

- For each $z \in \mathcal{Z}$, a vertex is present. It is colored by $g_{\mathcal{Z}}(z)$.
- An i-colored edge $(z_k, z_l) =: e_{k,l}$ is present if and only if there exists an input $i \in \mathcal{I}$ which maps z_k to z_l.

Note that the difference between a vertex which represents a cyclic attractor and a vertex which represents a fixed-point-attractor is that more than one line labelled with an input $i \in \mathcal{I}$ may leave the former, whereas exactly one line labeled with any $i \in \mathcal{I}$ leaves the latter.

An example of a mapping graph is depicted in Fig. 3.1. It is unlikely, however, that a mapping graph of such a simple nature appears in realistic situations.



	z_1	z_2	z_3
1	z_1	z_1	z_1
2	z_2	z_2	z_2
3	z_3	z_3	z_3

Fig. 3.1: Example of a mapping graph as defined in Sec. 3.1. The three vertices refer to the three attractors in $\mathcal{Z}$. Red and black circles around them encode the output of the attractors. The lines labeled $1, 2$ and 3 represent the attractor-correspondence, i.e. show what happens when an input $i \in \{1, 2, 3\}$ is presented to the NN in one of the three attractors. Note that the NN described by this mapping graph is a perfect learner (c.f. Sec. 1.3.3) and only possesses fixed-point-attractors.
For comparison, the attractor-correspondence of the NN is shown in the table. There, $f(i, z)$ is displayed for every input $i \in \{1, 2, 3\}$ and attractor $z \in \mathcal{Z} = \{z_1, z_2, z_3\}$.

Note that in the theory of finite-state automata, a similar concept is used, called "state diagram".

[2] A multigraph is a graph in which more than one edge can connect two vertices.

3.2 Attractors and histories

In this section, we seek a relation between the attractors of a NN and series of consecutively performed measurements. The latter have been defined as histories in Sec. 1.2 (p. 22).

The following questions will be answered in this section:

1. Given a series of measurements and respective results (i.e., a history h), can we specify which attractor the NN currently inhabits? Can one attractor $z \in Z$ be identified, or only a subset of $\mathcal{Z}$ consisting of several attractors?
2. Can we specify the attractor which a NN inhabits at a time t by making *subsequent* measurements?
3. Can we specify a minimal length κ of histories beyond which - independent of the measurements performed - a mapping from histories to individual attractors is possible? Put differently: Can we specify a number κ of measurements needed to know the individual attractor in which the NN settles with certainty?
4. Can we prepare a given NN in a specified attractor z?

All of those questions focus on the mapping between histories and attractors.

The essential assumption of this section is that the **mapping graph of the NN under investigation** with respect to a given set of inputs $\mathcal{I}$ **is known**. If this is not the case, Sec. 3.3 offers a procedure to obtain it - or a subgraph thereof - and thus allows for the concepts of this section to be applied to the NN under consideration.

In the following, we define the association of a history h, $a(h)$. Suppose we are given a k-history $h = ((i_1, o_1), (i_2, o_2), ..., (i_k, o_k))$, i.e. information about k measurements which have been performed on the NN: which input was shown and what was the respective result. $a(h)$ shall be defined such that it contains all attractors of $\mathcal{Z}$ which the NN could occupy once h has been observed.

In a first step, we identify all attractors of the NN whose output equals the output of the first measurement, i.e. o_1. Furthermore, we know that that the attractor which the NN occupies after the first measurement was reached by input i_1. Thus, all nodes of the mapping graph, which are colored by o_1 and for which a line i_1 exists which points towards them, represent attractors which the NN could possibly occupy. We define this set as $a_1(h)$.

In a second step, we repeat the same construction, i.e. we define $a_2(h)$ to consist of those attractors which are colored o_2 and have an i_2-labeled line pointing towards them. However, additionally, we know that the NN has to have settled in one of the attractors of $a_1(h)$ before input i_2 was presented. Thus, we can chose only those attractors of $\mathcal{Z}$ to be contained in $a_2(h)$ which have a line labeled i_2 leading *from* $a_1(h)$ *to them.*

In the following steps, we repeat this same construction: Every set $a_l(h)$ is defined to encompass those attractors $z \in \mathcal{Z}$ which have an output o_l and for whom an i_l-labeled line exists which

points from an attractor in $a_{l-1}(h)$ to z. The association of a k-history h (k is the length of the history) is thus given by $a_k(h)$, i.e. $a(h) := a_k(h)$. This is expressed mathematically in the following definition.

Association Let the association of a k-history h, $a(h)$, be a set containing those attractors which are compatible with h. Based on the mapping graph, $a(h)$ can be defined in the following way.

$$a_1(h) := \{z \in \mathcal{Z} \,|\, g_{\mathcal{Z}}(z) = o_1;\ \exists\, i_1\text{-labeled edge to } z\},$$

$$a_2(h) := \{z \in \mathcal{Z} \,|\, g_{\mathcal{Z}}(z) = o_2;\ \exists\, i_2\text{-labeled edge from } a_1(h) \text{ to } z\},$$

$$\ldots$$

$$a_k(h) := \{z \in \mathcal{Z} \,|\, g_{\mathcal{Z}}(z) = o_k;\ \exists\, i_k\text{-labeled edge from } a_{k-1}(h) \text{ to } z\}$$

$$a(h) := a_k(h).$$

The answer to question 1 posed above is thus: $a(h)$ comprises the attractors which the NN could occupy once the history h has been obtained.

Exactness We call the association of h **exact** if and only if it contains exactly one attractor, i.e.

$$a(h) \text{ exact } \Leftrightarrow |a(h)| = 1. \tag{3.1}$$

Thus, whenever the association of a history is exact, we have identified the one attractor which the NN occupies with certainty once this history is observed. I.e., h yields, in the sense relevant here, maximal information about the state of the NN.

The idea underlying $a(h)$ can also be extended in the following sense: Given a history h, which attractor of $\mathcal{Z}$ was present, say, after the l^{th} measurement of h ($l < k$ for a k-history h)? This question can be relevant, e.g., if other properties of the NN are studied simultaneously, c.f. Sec. 4.2.

The answer to this question is not given by $a_l(h)$ since the information about the measurements which have been performed subsequent to the l^{th} measurement decreases the number of possible attractors which the NN could have occupied after the l^{th} measurement, as illustrated in example 3.1.

Therefore, to answer the question we define the recursive association in the following. The idea behind the definition is to work backwards from the association of a k-history h in a manner which is similar to the definition of $a(h)$. E.g., in the first step we define $b_{k-1}(h)$ to encompass those attractors of $\mathcal{Z}$ which have output o_{k-1} and from which the NN can reach $a(h)$ if input i_k

is presented.

Let the **recursive association** $b_l(h)$ be the set of attractors which the NN could have occupied after the l^{th} measurement of a k-history h, i.e.

$$b_k(h) := a(h),$$

$$b_{k-1}(h) := \{z \in \mathcal{Z} \,|\, g_{\mathcal{Z}}(z) = o_{k-1};\ \exists\, i_k\text{-labeled edge from } z \text{ to } b_k(h)\},$$

$$b_{k-2}(h) := \{z \in \mathcal{Z} \,|\, g_{\mathcal{Z}}(z) = o_{k-2};\ \exists\, i_{k-1}\text{-labeled edge from } z \text{ to } b_{k-1}(h)\},$$

...

$$b_l(h) := \{z \in \mathcal{Z} \,|\, g_{\mathcal{Z}}(z) = o_l;\ \exists\, i_{l+1}\text{-labeled edge from } z \text{ to } b_{l+1}(h)\}.$$

The recursive association also allows us to consider the attractors which the NN could have occupied before the onset of measurements. If we denote by o_0 the output of the NN before a measurement was performed, then

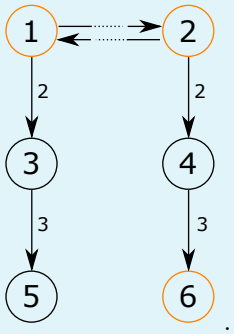
$$b_0(h) := \{z \in \mathcal{Z} \,|\, g_{\mathcal{Z}}(z) = o_0;\ \exists\, i_1\text{-labeled edge from } z \text{ to } b_1(h)\}$$

is a well-defined object. Thus, the answer to question 2 posed above is positive: Before measurements were carried out, the NN has certainly occupied one attractor of the set $b_0(h)$.

$a_l(h)$ and $b_l(h)$ may differ, as shown by the following example.

Example 3.1: Recursive association $b_l(h)$

In this example, we show that the association and the recursive association of a history h may differ. Assume we are dealing with a NN whose mapping graph has the following sub-structure (i.e., the following is a subgraph of the mapping graph of the NN), where the vertex-colours black (b) and orange (o) encode two different outputs of the respective attractor:



Suppose the history

$$h = \big((2, b), (3, o)\big)$$

has been obtained, then

$$\begin{aligned} a_1(h) &= \{3, 4\}, \\ a_2(h) &= \{6\} = a(h), \\ b_2(h) &= a(h) = a_2(h), \\ b_1(h) &= \{4\}, \end{aligned}$$

and thus $b_1(h) \neq a_1(h)$. This means that the additional measurement $(3, o)$ has led to an increase of information about the attractor after the first measurement.

We can easily determine whether two attractors z and z' of $\mathcal{Z}$ are distinguishable by measurements. Namely, if every history whose association contains one of them also contains the other, they cannot be distinguished by any number of measurements.

Indistinguishability Thus:

$$z, z' \in \mathcal{Z} \text{ indistinguishable } \Leftrightarrow \forall h : (z \in a(h) \Leftrightarrow z' \in a(h)), \tag{3.2}$$

which means that there is no history h whose association contains only one of the two.

Let the number κ mentioned in question 3 above, i.e. the minimal number such that every history of length $\geq \kappa$ can be associated with one individual attractor, be called the **historizon** of a NN.

To define it mathematically, we need the following two definitions. The first refers to the fact that there are some combinations of input and output which cannot be reproduced by a NN, e.g., if the NN never reaches an attractor with this particular output once this particular input is presented. The same holds for histories: There are histories which can be "written down" which cannot ever be reproduced by a NN. We call such histories impossible.

The second definition specifies whether every history of a NN with a particular length can be associated with an individual attractor of the NN. This is the case if the association of every history of that length, which is not impossible, contains exactly one attractor. Thus:

- Let a history h be **impossible** iff the given NN cannot reproduce it.
- Let the given NN be called k**-complete** iff all k-histories are either impossible or have exact associations.

Given those definitions, κ can be defined easily. It is given by the minimal length of histories for which a given NN is complete:

Historizon The historizon κ is given by

$$\kappa = k \;\Leftrightarrow\; \text{NN is } k\text{-complete but not } l\text{-complete } \forall\, l < k. \tag{3.3}$$

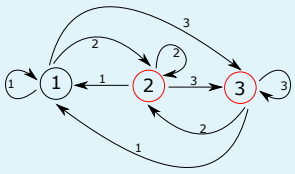
If a NN possesses two indistinguishable attractors, then, for every length l, there is a l-history which is possible and not exact. Thus, the NN is not l-complete $\forall\, l$. It follows that κ is undefined for this case, which is in accordance with its construction: there is no length of histories κ above which every history can be related to one individual attractor. If κ is undefined, question 3 has to be negated.

If a given NN has a finite historizon κ it is possible to prepare it in an arbitrary attractor $z \in \mathcal{Z}$ by showing the inputs of a history h whose association contains exactly z (question 4). Once the obtained outputs are exactly those specified in h, the attractor z is occupied with certainty. The history h exists (i.e. is realizable by the NN) and has length $\geq \kappa$ by construction of κ. If κ is undefined, question 4 has to be negated: the NN cannot be prepared in any $z \in \mathcal{Z}$.

Put differently, if defined, κ specifies how 'long' (in terms of measurements) an observer has to look back into the past to know the present attractor with certainty.

Example 3.2: A perfect learner

Suppose we are dealing with a NN whose behavior is captured by the mapping graph illustrated in Fig. (3.1):



We can then consider all 1-histories with input $\in \{1, 2, 3\}$ and output $\in \{\text{black (b), red (r)}\}$, namely: $(1, b)$, $(1, r)$, $(2, b)$, $(2, r)$, $(3, b)$, $(3, r)$. We immediately notice that $(1, r)$, $(2, b)$ and $(3, b)$ are not realizable by the network, thus impossible. For the others:

$$a\big((1, b)\big) = \{1\}$$
$$a\big((2, r)\big) = \{2\}$$
$$a\big((3, r)\big) = \{3\}.$$

Thus, this NN has historizon $\kappa = 1$.

Summary 6: Relation between attractors and histories

In Sec. 3.2, we have constructed mathematical tools which allow to use information gained from successive measurements to specify which attractor of $\mathcal{Z}$ the NN currently inhabits or has previously inhabited if the mapping graph of the NN under investigation is known (c.f. Sec. 3.3).

The **association** $a(h)$ (p. 53) allows to specify which attractors a NN can occupy once a history h of measurements has been obtained. The **recursive association** $b_l(h)$ (p. 54) allows to specify which attractors a NN can occupy after the l^{th} measurement of a history h. Both of those notions may be of interest if correlations of the NNs behavior with other properties or events are being studied, c.f. Sec. 4.2.

Furthermore, the historizon κ, defined on p. 56, specifies the minimal number of measurements which an observer has to carry out in order to know the attractor of a NN with certainty. It also specifies the minimal number of measurements needed in order to prepare a NN in an attractor of $\mathcal{Z}$.

3.3 Reconstruction of the mapping graph

In this section we construct a procedure which allows to obtain the mapping graph (or a well-defined subgraph thereof) from a NN under investigation. The only means of investigations are measurements as presupposed throughout this thesis.

The following questions will be answered:

1. Given a NN in an experimental situation *about which no information is available*. How can a sensible study based on measurements be performed? Which information can be extracted and how is best represented?
2. Is there a criterion which specifies that all information which can be obtained has been obtained by the proposed means of investigation?

Ensemble In order to answer those questions, we make one major assumption: We assume that there is a suitably large ensemble of NNs - all in the same attractor - at the inquirer's disposal. Even tough this is usually not fulfilled in realistic situations, the existence of an ensemble can be mimicked if it is possible to prepare the NN in the same attractor again and again. That this is in in principle possible, and how it can be done, is mentioned in Sec. 3.2 (p. 56). E.g., if the NN always runs into the same attractor z upon presentation of input i, the NN can be prepared in this attractor by showing the input i, c.f. Sec. 4.1. The procedure can also be applied if the NN under investigation is a perfect learner with respect to at one least input (i.e., if Eq. (1.32) holds for at least one input i).

The availability of an ensemble allows us to deal with what we can call a full-k-set of histories, defined as follows:

Full-k-set of histories Recall the idea behind a history: We start with the NN in an (unknown) attractor $z \in \mathcal{Z}$ with (known) output $o_0 = g_{\mathcal{Z}}(z)$. A measurement is carried out by showing an input i_1 and obtaining, in the sequel, the respective output o_1, which will be one out of the set $\{g_{\mathcal{Z}}(f(i, z))\}$.[3] We repeat this step k times to obtain a k-history $h = ((i_1, o_1), ..., (i_k, o_k))$. Since an ensemble of NNs occupying attractor z is available, we can take a second NN and apply, as a first measurement, input j_1. Of course we are free to choose inputs as we like in the following, constituting a different k-history $h' = ((j_1, o'_1), ..., (j_k, o'_k))$.

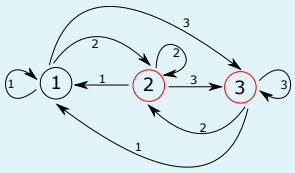
We define a **full-k-set of histories** to be a set of histories $H = \{h_1, h_2, ...\}$, where for all possible orderings of k elements of $\mathcal{I}$ (repetitions allowed) a history is contained in H which consists of exactly this ordering of measurements.

Put differently, a full-k-set of histories is obtained in the following way. First, divide your ensemble into $|\mathcal{I}|$ disjoint subsets and perform a measurement associated with each of the inputs $i \in \mathcal{I}$ on each of the subsets. Next, divide each of the subsets into $|\mathcal{I}|$ subsets and again show all of the inputs of $\mathcal{I}$ to one of the subsets. Repeat this for k steps.

The following example illustrates the definition.

Example 3.3: Full-2-set of histories

A full-2-set of histories of the perfect learner given by the mapping graph



is

$$H = \big\{\big((1,b),(1,b)\big), \big((1,b),(2,r)\big), \big((1,b),(3,r)\big)$$
$$\big((2,r),(1,b)\big), \big((2,r),(2,r)\big), \big((2,r),(3,r)\big)$$
$$\big((3,r),(1,b)\big), \big((3,r),(2,r)\big), \big((3,r),(3,r)\big)\big\},$$

where of course the redundant form of the full-2-set is due to the 'simple' nature of the NN.

In the following, we aim to develop an algorithmic procedure to extract information from the NN, which

[3] Note that here and in the following, we chose the same notational convention as in Eq. (1.32): We write $g_{\mathcal{Z}}(\{a, b, c\})$ instead $\{g_{\mathcal{Z}}(a), g_{\mathcal{Z}}(b), g_{\mathcal{Z}}(c)\}$.

a) yields a subgraph of the mapping graph of the NN under investigation.[4] This means that no non-existing attractors shall be represented in the constructed graph.

b) is guaranteed to converge, i.e. guaranteed to terminate.

Ontic mapping graph (omg) We refer to the 'true' mapping graph of the NN under investigation as the ontic mapping graph.

Output-structure We consider two vertices z and z' of a mapping graph to have the same output-structure if for each line labelled i leaving z towards a vertex colored $g_{\mathcal{Z}}$ there exists exactly one line i leaving z' towards a vertex colored $g_{\mathcal{Z}}$. Formally:

$$\begin{aligned} &z,\, z' \text{ have the same output-structure} \Leftrightarrow \\ &\forall\, i \in \mathcal{I}_z \,:\, \Big(g_{\mathcal{Z}}(f(i,z)) = \{o'_1, o'_2, .., o'_m\} \Leftrightarrow g_{\mathcal{Z}}(f(i,z')) = \{o'_1, o'_2, .., o'_m\}\Big), \end{aligned} \tag{3.4}$$

where $\mathcal{I}_z$ denotes the set of lines $i \in \mathcal{I}$ leaving z and where $o'_1, o'_2, .., o'_m \in \mathcal{G}_{\mathcal{Z}}$.

In the procedure which follows, we will use data obtained from the NN by measurements to construct a subgraph $\mathcal{M}'$ of the omg. Every vertex in $\mathcal{M}'$ will represent a vertex in the omg and thus an attractor $z \in \mathcal{Z}$. During the procedure, we will use the data from measurements to determine the output-structure of each vertex. It is essential that if cyclic attractors are present, the same vertex may have different output-structures at different times. The reason for this is that the output-structure depends on the *activation* (p. 8) which the NN occupies when the inputs are presented.

3.3.1 The reconstruction procedure $\mathcal{P}$

In the following, let the index l denote to the different histories present in a set of histories and carry this notation into the elements of the histories as

$$h_l =: ((i_{l,1}, o_{l,1}), ..., (i_{l,k}, o_{l,k})).$$

In the following procedure, a graph will be constructed successively, which we chose denote by $\mathcal{M}'$. It has already been mentioned that it is a subgraph of the mapping graph (but c.f. Sec. 3.3.3), hence formally it is also a directed edge- and vertex-colored multigraph.

In order to specify the procedure, we will have to refer to particular vertices of $\mathcal{M}'$. We do this in an unconventional but simple way. If, e.g., a vertex with color $o_{l,m}$ has been inserted in $\mathcal{M}'$, we refer to this vertex subsequently by "$o_{l,m}$".

Note that o_0 as been defined on p. 58 in the context of a full-k-set of histories.

[4] Note that a graph is its own subgraph, c.f. e.g. [Har11].

The reconstruction procedure $\mathcal{P}$

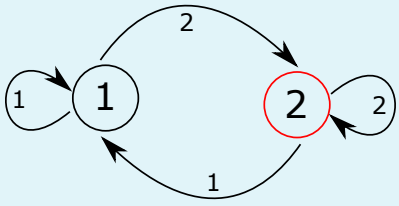
1. Draw a vertex which is colored o_0 into $\mathcal{M}'$.
2. Obtain the full-1-set of histories.
3. For each element of the full-1-set, draw a vertex colored with the obtained output, $o_{l,1}$, and insert a line from the o_0-vertex to the $o_{l,1}$-vertex into $\mathcal{M}'$. Label the line $i_{l,1}$.
   $m := 2$.
4. Obtain the full-m-set of histories. Specify which output-structure belongs to each vertex in $\mathcal{M}'$. For cyclic attractors, this may be a different one from the one it already has (see above). In this case, both (all) specified output-structures belong to the vertex.
5. Merge (combine) vertices of $\mathcal{M}'$ which have at least one output-structure in common *and* the same output-value.
6. Draw the newly obtained information into $\mathcal{M}'$: If there is *no* line labeled $i_{l,m}$ leaving the vertex $o_{l,m-1}$ (or the one it has been merged into) towards a vertex colored $o_{l,m}$, create a new vertex colored $o_{l,m}$ and insert the line. If there is a line labeled $i_{l,m}$ leaving the vertex $o_{l,m-1}$ (or the one it has been merged into) towards a vertex colored $o_{l,m}$, change nothing but refer to the latter by $o_{l,m}$.
7. $m := m + 1$ and continue with step [4.] unless $\mathcal{M}'$ has not changed in the last X steps, where X is an (in general small) integer to be specified by the inquirer. The relevance of X is specified below (Sec. 3.3.4).

We refer to this procedure by $\mathcal{P}$.

In the following, we will consider two explicit examples of the procedure, one consisting of fixed-point-attractors only, and one including a cyclic attractor. In Sec. 3.3.2, we will prove the termination of $\mathcal{P}$ for NNs comprising fixed-point-attractors only, and in Sec. 3.3.3, we do the same for NNs comprising also cyclic attractors. In Sec. 3.3.4 we will comment on the constraints a) and b) mentioned above and add some notes.

Example 3.4: A simple NN

This example demonstrates $\mathcal{P}$ for the rather simple NN given by the omg



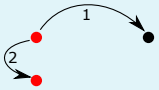
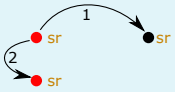
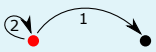
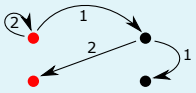
We assume an ensemble of such NNs is given and that all of its members occupy attractor

$z = 2$. Equivalently, we can assume that one NN is given since this NN can be prepared in attractor 1 an arbitrary number of times by showing input 1.

In a realistic scenario, neither the omg, nor z would be known to an observer. However, since we want to demonstrate how $\mathcal{P}$ functions, we will use the knowledge of the omg and of z to generate the full-k-sets of histories which would otherwise be obtained through measurements. Using the those, we will apply $\mathcal{P}$ and indeed recover the omg.

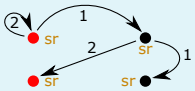
Let $k_{m=l}$ denote step k of $\mathcal{P}$ with $m = l$. We include all relevant details in this example.

$\mathcal{P}$ yields:

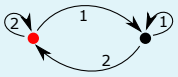
Step(s)	$\mathcal{G}$
1	
2	Full-1-set: $\{((1,b)),\ ((2,r))\}$, where b=black and r=red
3	
$4_{m=2}$	Full-2-set: $\{((1,b),(1,b)),\ ((1,b),(2,r)),\ ((2,r),(1,b)),\ ((2,r),(2,r))\}$ Output-structure: (sr = (1,b),(2,r)) 
$5_{m=2}$	
$6_{m=2}$	

$4_{m=3}$ Full-3-set: $\{((1,b),(1,b),(1,b)),\ ((1,b),(1,b),(2,r)),$
$((1,b),(2,r),(1,b)),\ ((1,b),(2,r),(2,r)),\ ((2,r),(1,b),(1,b)),$
$((2,r),(1,b),(2,r)),\ ((2,r),(2,r),(1,b)),\ ((2,r),(2,r),(2,r))\}$

Output-structure:

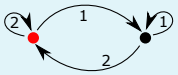


$5_{m=3}$



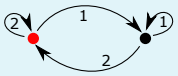
No new lines are created.

$6_{m=3}$



This is true for all future steps, hence $\mathcal{P}$ terminates.

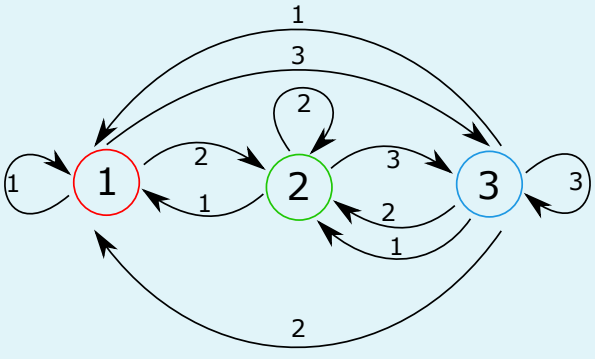
$4_{m=4}$, $5_{m=4}$, ...



Thus, the omg is recovered by $\mathcal{P}$. A more involved case is discussed in example 3.5.

Example 3.5: Cyclic attractors

In this example we demonstrate $\mathcal{P}$ on an ensemble of NNs given by the omg



,

where all NNs of the ensemble shall occupy attractor 1. We emphasize again that we use the information represented by the omg only to generate the full-k-sets of histories which would usually be obtained through measurements.

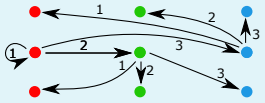
Attractor 3 is a cyclic attractor. This can be inferred from the fact that two lines labeled with the same input are leaving it. Therefore, whenever a NN occupies attractor 3, the output-structure is either rgb (i.e., $(1,r)$, $(2,g)$, $(3,b)$) or grb (i.e., $(1,g)$, $(2,r)$, $(3,b)$), where r, g and b (red, green and blue) represent the three different outputs of the NN. Which of those two possibilities is the case depends on which activation of the NN is occupied when the inputs are shown. Since this cannot be known to an inquirer (c.f. footnote 1 on p. 49), either of the two possibilities can occur with some probability.

We choose the same notation as in example 3.4 but state the full-k-sets of histories only for small k because they quickly become large.

$\mathcal{P}$ yields:

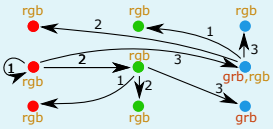
Step(s)	$\mathcal{G}$
1	
2	Full-1-set: $\{((1,r)),\ ((2,g)),\ ((3,b))\}$, where r=red, g=green and b=blue
3	
$4_{m=2}$	Full-2-set: $\{((1,r),(1,r)),\ ((1,r)(2,g)),\ ((1,r),(3,b)),$ $((2,g),(1,r)),\ ((2,g),(2,g)),\ ((2,g),(3,b)),$ $((3,b),(1,g)),\ ((3,b),(2,r)),\ ((3,b),(3,b))\}$ Output-structure: (rgb=(1,r),(2,g),(3,b); grb=(1,g),(2,r),(3,b)) 
$5_{m=2}$	

$6_{m=2}$

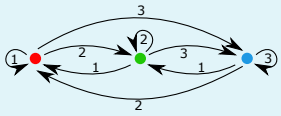


Output-structure:

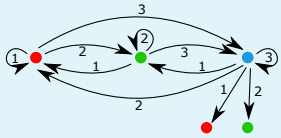
$4_{m=3}$



$5_{m=3}$

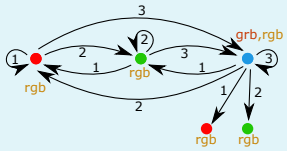


$6_{m=3}$

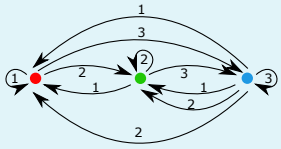


Output-structure:

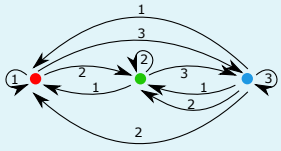
$4_{m=4}$



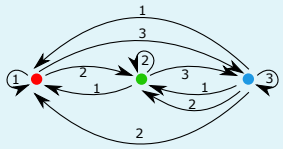
$5_{m=4}$



$6_{m=4}$



$4_{m=5}$, $5_{m=5}$, ...



No new lines are created, thus $\mathcal{P}$ terminates.

3.3.2 Proof of termination for NNs comprising fixed-point-attractors only

Since the the termination of $\mathcal{P}$ can be proven in a conceptually clearer way if it is assumed that the NN on which $\mathcal{P}$ operates consists of fixed-point-attractors only, we will do so in this section. The more general case is presented in Sec. 3.3.3.

Lemma 3.1. *$\mathcal{P}$ terminates for NNs comprising fixed-point-attractors only.*

Proof We have mentioned before that $\mathcal{Z}$ is finite (p. 15). Additionally, we have noted that the attractor-correspondence is not probabilistic for fixed-point attractors (p. 18). Thus, whenever all inputs of the an input set $\mathcal{I} \subseteq \mathcal{I}^*$ are shown to NNs of an ensemble in an attractor z, as is the case when a full set of histories is obtained by measurements on this ensemble, the NNs of the ensemble will subsequently occupy all attractors which can be mapped to from z.

Since $\mathcal{Z}$ is finite and closed (p. 15) this implies that within a finite length of full-history-measurements, say after a full-k-set of histories has been obtained, all attractors that can be reached by NNs of the ensemble will have been reached by the NNs of the ensemble. $(*)$

We say that a line $e_{k,l}$ of the mapping graph underlying a NN, which connects the vertices z_k and z_k and carries a label i, "has been used" once the NN has made the transition to z_l after the input i was shown to the NN inhabiting attractor z_k. The finiteness of $\mathcal{I}^*$ and $(*)$ imply that after the full-$(k+1)$-set of histories has been obtained, all lines of the omg underlying the ensemble have been used at least once.

Due to the merging occurring in step 5 of $\mathcal{P}$, no two vertices of $\mathcal{M}'$ can represent the same fixed-point attractor. Thus, in step 6 of $\mathcal{P}$, new lines can only be added to $\mathcal{M}'$ if a NN of the ensemble uses a line of the omg which has not been used before by any NN of the ensemble. Thus, whenever all lines of the omg underlying the NNs of an ensemble have been used, as is the case for $m \geq k+1$, no new lines will be added to $\mathcal{M}'$.

Thus, $\mathcal{M}'$ will not change for $m > (k+1)$, which implies that $\mathcal{M}'$ converges and $\mathcal{P}$ terminates.

Q.e.d.

3.3.3 Proof of termination for general NNs

Due to the probabilistic behavior of cyclic attractors, the finiteness of $\mathcal{Z}$ and $\mathcal{I}$ does not suffice to guarantee that all lines which can be used by the NNs of an ensemble have been used after a finite length of full-history-measurements. However, the following lemma implies the termination of $\mathcal{P}$. Let X and m be the integers defined in $\mathcal{P}$ above. m mainly controls the length of full-sets of histories which have been obtained.

Lemma 3.1. $\exists\, l$: *for* m *with* $l \leq m \leq l + X$ *no lines of the omg are used which have not been used before.*

Proof The proof uses reductio ad absurdum. Suppose the negation of lemma 3.1 holds, i.e.:

$$\forall\, l \;:\; \text{for } m \text{ with } l \leq m \leq l + X \text{ lines of the omg are used which have not been used before.} \tag{3.5}$$

Since there is no upper bound on l, it follows from statement (3.5) that the number of lines in the omg is arbitrarily large. This contradicts the finiteness of $\mathcal{Z}$ and $\mathcal{I}$. Therefore, (3.5) is false and lemma 3.1 is true. *Q.e.d*

Lemma 3.1 implies the following theorem.

Theorem 2

$\mathcal{P}$ terminates for any NN compatible with the definitions of Ch. 1.

Proof As we have shown in the proof of lemma 3.1, if no lines of the omg are used which have not been used before within a round of $\mathcal{P}$, no new lines are added to $\mathcal{M}'$. Lemma 3.1 shows that there exits a l such that for m with $l \leq m \leq l + X$, no new lines of the omg are used, hence $\mathcal{M}'$ does not change for X steps, hence $\mathcal{P}$ terminates. *Q.e.d.*

Relevance of X The last proof exposes the relevance of X. The larger X is, the higher is the probability to use all lines of the omg leaving a cyclic attractor at least once, and thus to find them included in $\mathcal{M}'$. That is to say, the larger X, the higher the probability to show the inputs at least once to every activation constituting a cyclic attractor.

3.3.4 Comments

In this section, we elaborate on the difference between $\mathcal{M}'$, which is constructed successively by $\mathcal{P}$, and the mapping graph of the NN under investigation (i.e., its omg), answer the questions posed on p. 57, and make a comment about the connection between Secs. 3.2 and 3.3.

$\mathcal{M}'$ is a subgraph of the omg of the NN under investigation. It differs from the latter in two ways:

First, the merging which takes place in step 5 of $\mathcal{P}$ suppresses some of the omg-vertices. Namely, whenever two attractors z and z' have

a) the same output-value, i.e. $g_{\mathcal{Z}}(z) = g_{\mathcal{Z}}(z')$ and

b) at least one output-structure in common (as defined on p. 59),

they can *not* be distinguished by $\mathcal{P}$. Thus $\mathcal{M}$ represents such attractors in one vertex.

Second, since the behavior of cyclic attractors is probabilistic, it is possible that some of the omg-lines of cyclic attractors are not visible in $\mathcal{M}'$. This can be modulated by X, as defined in $\mathcal{P}$. The larger X, the larger the probability to find all of the omg-lines of cyclic attractors in $\mathcal{M}'$.

On p. 57, we have posed two questions which guided the investigation of Sec. 3.3:

1. Given a NN in an experimental situation *about which no information is available*. How can a sensible study based on measurements be performed? Which information can be extracted and how is best represented?

2. Is there a criterion which specifies that all information which can be obtained has been obtained by the proposed means of investigation?

We have answered question 1 by constructing a procedure $\mathcal{P}$ which allows to obtain the mapping graph of a NN - or a subgraph thereof, c.f. above - from a NN under investigation by performing measurements. The mapping graph of a NN contains information about the attractors of NN, the behavior of the NN once input is presented and the output of a NN. Thus, it represents information about the NN under investigation in a sensible way.

Question 2 has been answered by the construction of $\mathcal{P}$, which contains a sensible termination-condition. This has been proven based on the definitions of Ch. 1, i.e., it holds for all NN which are compatible with them.

Once $\mathcal{M}'$ has been obtained for a given NN, the methods of Sec. 3.2 can be used for further studies. In this case, the just-mentioned difference between $\mathcal{M}'$ and the omg of the NN under investigation carries over to the results of Sec. 3.2. E.g., in this case, the results of Sec. 3.2 cannot discriminate between attractors which have a) and b), as mentioned above, in common.

Ch. 4 further elaborates on this point and outlines possible applications.

Summary 7: A procedure to reconstruct the mapping graph of a NN through measurements

In Sec. 3.3 we have developed a procedure $\mathcal{P}$ (p. 59) which allows to reconstruct the mapping graph (c.f. Sec. 3.1) of a NN (or a subgraph thereof, c.f. Sec. 3.3.4) through measurements on those NNs (c.f. Sec. 1.2).

Several examples in this section serve to illustrate $\mathcal{P}$.

3.4 Feedforward NNs

A feedforward NNs is a type of NN which has a very simple structure. E.g., information can only travel in one direction in a feedforward NN. For this reason, it offers itself as a suitable example to illustrate some concepts of this thesis, which constitutes the scope of this section.

A feedforward NN can be defined in the following way. Suppose the neurons of a NN are arranged in k layers, as illustrated in Fig. 3.2. Suppose furthermore that the first layer serves as input of the NN and that the k^{th} layer serves as output of the NN. If the edges of the graph $\mathcal{M}$ associated with a NN (c.f. p. 6) obey the following requirement, this NN is called a **feedforward NN**:

(fN) The neurons of the i^{th} layer are only connected to the neurons of the $(i+1)^{th}$ layer, where $i = 1, ..., k-1$. I.e.: Each vertex of $\mathcal{M}$ in layer i has only directed lines leaving it, which point towards a vertex in layer $i+1$. No connections within one layer, no "backwards"-connections and no cycles exist.

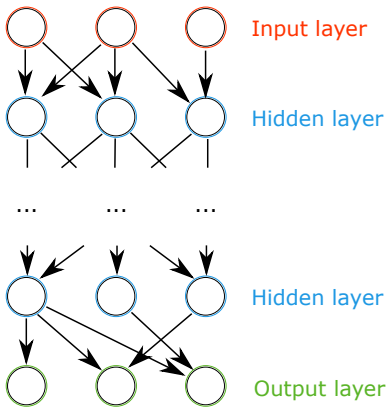


Fig. 3.2: Schematic illustration of a feedforward NN.

This suffices to guarantee that a feedforward NN only comprises fixed-point attractors, which can be understood in the following way. If an input is presented to the NN, the neurons in layer 1 are fixed. Hence, the neurons in layer 2 have to be stable after some time (how long exactly depends on the update-rule which is assumed), hence the neurons in layer 3 have to be stable after some time, and so forth. Therefore, the attractor into which the NN runs is a fixed-point attractor. This implies that the attractor-correspondence f, as introduced in Sec. 1.1.5, is a

function (c.f. p. 17).

The information in a feedforward NN can only travel in one direction. This follows from requirement (fN) above since the neurons of a layer j cannot influence the neurons of a layer i with $i < j$. This, in turn, implies that the behavior of the NN after an input is presented is independent of the states of all but the input-neurons. Put differently: The attractor, in which a feedforward NN runs is only dependent on the input, which is presented, and not on any previous state of the NN. Therefore, any feedforward NN obeys

$$f(i, z_k) = f(i, z_l) \qquad \begin{array}{l} \forall\, i \in \mathcal{I}, \\ \forall\, z_k, z_l \in \mathcal{Z}, \end{array} \tag{3.6}$$

which implies immediately that it obeys the perfect-learner condition (1.32).

With this information, we can draw the mapping graph of any given feedforward NN immediately. Suppose that the number of inputs, which is specified, is three, i.e.: $\mathcal{I} = \{1, 2, 3\}$. If all inputs are recognized separately by the NN (this means that they all lead to a different output), this implies that $\mathcal{Z} = \{z_1, z_2, z_3\}$. Thus, the mapping graph of any feedforward NN is given by Fig. 3.3. This generalizes to any number of inputs.

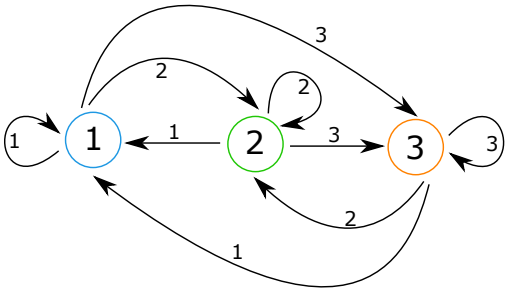


Fig. 3.3: Illustration of the mapping graph of a feedforward NN for $|\mathcal{I}| = 3$. The colors of the vertices represent the output of the NN.

This demonstrates that the mapping graph of a NN, which is necessary for the concepts of Sec. 3.2, can be sometimes be obtained by simple thoughts.

Once a mapping graph is given, the concepts of Sec. 3.2 can be applied. In example 3.2, this is done for a NN whose a mapping graph is very similar to the one of Fig. 3.3. However, in case a feedforward NN is given, it is clear that the NN can be prepared in any one of its attractors by showing one input only. This implies that the historizon κ of a feedforward NN is 1: $\kappa = 1$.

We conclude this section with a quick reference to chapter 2. Eq. (3.6) implies that

$$f(i, f(j, z)) \neq f(j, f(i, z)).$$

With Eqs. (2.18) and (2.8), this implies that the observables of feedforward NNs do *not* commute. Since feedforward NNs obey the perfect learner condition (1.32), this is in accordance with lemma 2.1, whose prove utilizes perfect learners.

Summary 8: Feedforward NNs

In this section, we study feedforward NNs (c.f. Fig. 3.2), which serve to illustrate some results of this thesis.

Feedforward NNs have fixed-point-attractors, hence the attractor-correspondence is function. Since information can only travel in one direction, they satisfy the perfect learner condition (1.32). This allows to construct the mapping graph of a feedforward NN. If $\mathcal{I}$ is specified to contain three inputs, the mapping graph of a feedforward NN is given in Fig. 3.3.

The observables of feedforward NNs do *not* commute.

Chapter 4

Remarks and applications

In Ch. 3, we have developed mathematical tools which allow to identify attractors of a NN through series of consecutively performed measurements (histories). The motivation for this inquiry, which has been outlined from p. 49 onwards, is of a particular theoretic nature.

Yet, the notion of measurement, which underlies all considerations of this thesis, is explicitly motivated by experiments on NNs. Therefore, in this chapter, we will make some remarks about whether concepts of Ch. 3 may be of interest in actual experiments. This will be complemented by remarks about how to apply those concepts in experiments. During all of this, we put an emphasis on explanations which are not abstract or mathematical in nature.

We emphasize that Ch. 3 assumes, as all other chapters so far, that the NNs, which are examined by measurements, are compatible with the definitions of Ch. 1. We have noted before (p. 18, 3) that even though this last requirement encompasses a multitude of different NNs, not all NNs are compatible with those definitions. Particularly, it rather is unlikely that networks of biological neurons obey the definitions.

This implies that with respect to biological NNs, the definitions of Ch. 1 can only be understood to constitute a toy-model, i.e. a model which is simplified compared to biological networks of neurons, but complex enough to show non-trivial properties.

Contrary to that, that the notion of 'output of a measurement' is not limited to what is conventionally understood as 'output' of a NN. Rather, we have only required the relation between a NN's activation and the output of a measurement to be a function (c.f. Sec. 1.1.2). Therefore, the definitions of Ch. 1 also capture scenarios in which a NN is investigated by use of a measurement-device. This is illustrated in example 1.2 (p. 9).

In Sec. 4.1, we will offer a hands-on illustration of how the mapping graph reconstruction procedure $\mathcal{P}$, as introduced in Sec. 3.3, works. In Sec. 4.2, we will do the same for the concepts 'association' and 'recursive association' which have been introduced in Sec. 3.2. In Sec. 4.3, we will remark about the difference between those three concepts and a conventional mathematical analysis of measurements.

4.1 Remarks about the mapping graph reconstruction procedure $\mathcal{P}$

The mathematical tools developed in Sec. 3.3 allow to study the behavior of a NN without knowing anything about it explicitly. No knowledge about the type of neurons present, their possible states, the topology of the NN, the synaptic transmissions or the dependency of the NN's output on the states of the neurons is necessary.

The procedure $\mathcal{P}$ allows to extract information about the NN and represent it in terms of the mapping graph, which has been introduced in Sec. 3.1. As noted in Sec. 3.3 (p. 57), $\mathcal{P}$ is can be applied to a single NN e.g. if the NN behaves be a perfect learner (introduced in Sec. 1.3.3) with respect to at least one input, say i_p. The following example outlines an application of $\mathcal{P}$ if the latter is assumed.

Assume you are an experimenter who wishes to study the structure of a NN without destroying it (compare in-vivo studies in biology or neuroscience) and that in addition, you know nothing about the NN besides the fact that it obeys the conditions imposed in Ch. 1.

Before starting your study, choose a suitable class of inputs $\mathcal{I}$. Next, present i_p to the NN to make sure that it is in a state into which you can bring it back at any time. Register the output of your measurement-device before you make further measurements. This is o_0 used in the first step of $\mathcal{P}$.

In order to apply $\mathcal{P}$, you have to obtain several k-histories, i.e. sets of k measurements, from the NN. To obtain one single k-history, decide for a combination of k inputs from $\mathcal{I}$ and perform the measurements: Show the first input, wait until the NN has settled into an attractor (c.f. Sec. 1.1.3.1) and read off the output of your measurement-device. Then show the next input of your combination and so forth.

After you have shown all k inputs, obtain a different k-history: Bring the NN back into the starting-state by showing i_p, choose a different combination of k inputs and obtain the respective k-history.

Repeat those steps until you have obtained a full-k-set of histories, which is the case if you have shown all possible combinations of k inputs (repetitions allowed) to the NN and obtained the associated histories.

Once this is done for a sufficiently high k,[1] you have all the information required to use $\mathcal{P}$ to construct a graph $\mathcal{M}'$, as described in Sec. 3.3.1. $\mathcal{M}'$ is a subgraph of the mapping graph of the NN under investigation, i.e., you have acquired information about the behavior of the NN: Which attractors exist, how the NN behaves when an input is presented and which output the attractors have.

[1] Note that a full-l-set of histories is contained in a full-k-set if $l < k$. Thus, all the full sets of histories required for the individual steps of $\mathcal{P}$ can be read of a full-k-set with sufficiently large k.

Thus, $\mathcal{P}$ can be used to investigate the input-dependent behavior of a NN. The limitations of $\mathcal{P}$ concerning attractors, which cannot be distinguished, are explained in Sec. 3.3.4.

We emphasize that no form of statistical averaging is contained in $\mathcal{P}$, which may be useful in some situations.

4.2 Remarks about the association $a(h)$ and the recursive association $b_l(h)$

The concepts of Sec. 3.2 allow to identify which attractors a NN inhabits after (or before) one or more measurements have been performed. This is of importance if several attractors have the same output, which will likely be the case if only a small part of the neuron-states determines the output of a measurement, as is e.g. the case in single-unit-recordings in biology.

In order to use the concepts, some information about the NN is required: the mapping graph of the NN under investigation. Since a subgraph of the mapping graph can be obtained through $\mathcal{P}$, the concepts of Sec. 3.2 can also be applied to a NN about which no information is available. In this case, $\mathcal{P}$ has to be applied previous to the concepts of Sec. 3.2.

A history is a series of consecutive measurements (c.f. Sec. 1.2.3). Once the mapping graph is available, the association of a history h, $a(h)$, specifies which attractors the NN could occupy after this history has been obtained. The recursive association $b_l(h)$ specifies which attractors could have been occupied by the NN after the l^{th} measurement of this history, and $b_0(h)$ specifies which attractors could have been occupied by the NN directly before the measurements began.

In order to use either $a(h)$ or $b_l(h)$, a history has to be obtained from the NN, i.e. a sequence of measurements has to be carried out, nothing more. The mathematics of the respective definition immediately give the desired result.

Thus, the concepts of Sec. 3.2 and 3.3 together allow to identify the mapping-graph of a NN and, subsequently, allow to use this information to constrain which attractors a NN may inhabit before or after any measurement. I.e., for NNs compatible with Ch. 1, they form a set of tools which allows to analyze the NN under investigation in an unconventional way.

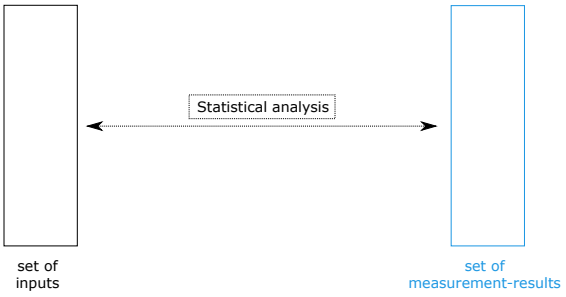
4.3 Comparison with a conventional analysis

In this section, we will compare the concepts of Ch. 3 to a conventional analysis of 'measurements'. However, since we do not take into account any details about particular experimental setups, this comparison remains preliminary.

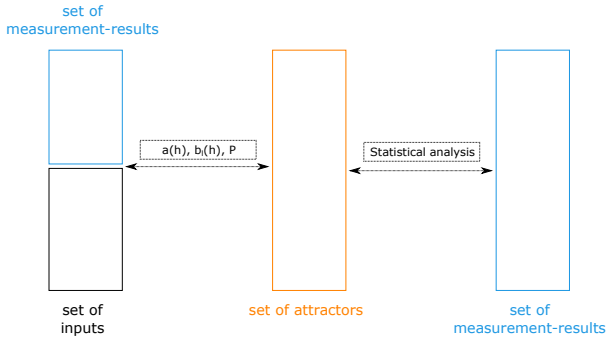
Essential about the concepts of Ch. 3 is that they allow to infer the closed set of attractors of a NN for a given input-set $\mathcal{I}$, and that they furthermore allow to determine which of those

attractors is occupied by a NN after or before measurements on this NN have been performed.

Thus, whenever the input-dependent behavior of a NN is investigated, e.g. by statistically analyzing the relation between the input, which has been shown, and the respective output of the NN (as illustrated in Fig. 4.1(a)), the concepts of Ch. 3 open the possibility for a statistical analysis of the relation between the attractor, which a NN has occupied after a measurement, and the respective output of the NN (c.f. Fig. 4.1(b)).



(a)



(b)

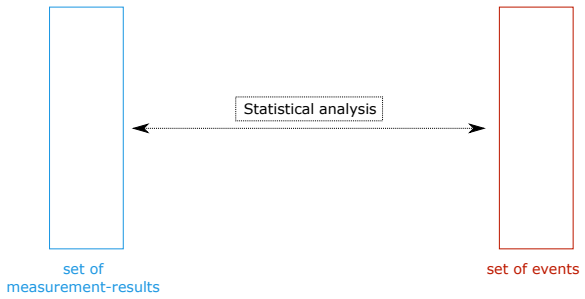
Fig. 4.1: Schematic illustration of the difference between a conventional analysis of measurements and an analysis which utilizes $\mathcal{P}$, $a(h)$ and $b_l(h)$, which have been introduced in Ch. 3. C.f. Sec. 4.3.

The same holds if the correlation between the output of measurements and other events is studied (Fig. 4.2(a)). E.g., the output of a NN could be compared with motor-events (biology), states of consciousness (neuroscience) or stock-market behavior (economics). Again, the set of attractors can be obtained together with information about which attractor was present after individual measurements, and a statistical analysis of the relation between this information and the set of events can be performed (Fig. 4.2(b)), which might yield additional insights into the system under investigation.

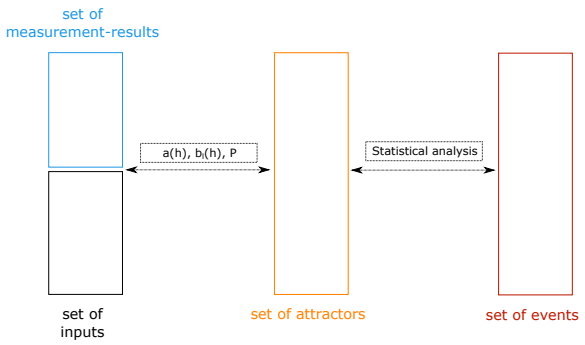
Thus, the main difference between a conventional analysis of measurements and an analysis uti-

lizing $\mathcal{P}$, $a(h)$ and $b_l(h)$ is that the latter offer information about the attractor of the NN, which was occupied at a certain time. This information could be used to gain additional insights into the NN under investigation.

We conclude this chapter with a remark that none of the concepts of Ch. 3 include any form of statistical averaging. This might offer a possibility to study NNs completely without use of the latter, which would certainly yield new insights in some situations.



(a)



(b)

Fig. 4.2: Schematic illustration of the difference between a conventional analysis of the relation between measurements and events, which can be associated with a NN, and an analysis which utilizes $\mathcal{P}$, $a(h)$ and $b_l(h)$. C.f. Sec. 4.3.

Summary 9: Remarks and applications

Ch. 4 offers further remarks about the concepts which have been developed in Ch. 3 and elaborates on how they are applied. (Sects. 4.1 and 4.2)

In Sec. 4.3, an inquiry based on $\mathcal{P}$, $a(h)$ and $b_l(h)$ is compared with a direct statistical

analysis of a) the relation between input and measurement-results and b) the relation between measurement-results and other events. In both cases, the information about the attractors of a NN, which can be assessed through $\mathcal{P}$, $a(h)$ and $b_l(h)$, offers an additional perspective on the system's dynamics. It can be compared with measurement-results or events in a statistical way.

No form of statistical averaging is included in $\mathcal{P}$, $a(h)$ or $b_l(h)$ (p. 75). Thus, those concepts can be used to study a NN under investigation without utilizing statistical averaging at all, which might yield new insights in some situations.

Chapter 5

CHSH inequalities

In a thesis about the mathematical structure of measurements, observables and states on NNs one would not expect a chapter about CHSH-inequalities. The reason of why such a chapter is included in this thesis is twofold, as will be explained in detail in Sec. 5.1.

One reason is related to the fact that observables on NNs, as elaborated in Ch. 2, do not commute, which in turn is a consequence of our definition of measurement. Non-commuting observables are, in general, not expected in the realm of classical physics and usually considered to be a property of quantum systems. (In contrast, non-commuting processes are abundant in classical physics as well.) Therefore, it is a legitimate question of whether quantum-like phenomena are to be expected in this class of models. Results of this direction have been reported in [bGFA12, bGA06, Fil12].

The second reason of why a chapter about CHSH-inequalities is included here relates to findings in the cognitive sciences. Several recent articles report about psychological experiments in decision making [BB12], cognition [KC08] and language usage [ACD06], where a violation of CHSH-inequalities, or other Bell-type inequalities, is observed. The question arises under which conditions such data can be explained in classical systems. The investigation of NNs may be of help in this context.

In Sec. 5.1, a detailed account of the motivation driving this chapter is given. In Sec. 5.2, an introduction to the CHSH-inequality is offered. It anticipates some results of Sec. 5.3, where a formal derivation of the CHSH-inequality is given. Sec. 5.4 elaborates on the logical implications of a violation of the CHSH-inequality for arbitrary systems. Finally, Sec. 5.5 explains why and proves that the CHSH-inequalities can be violated by NNs.

Note that throughout this chapter, no space-like separation of measurements is presupposed. This is in accordance with all references of this chapter which associate Bell-type inequalities with non-quantum-theoretic systems, i.e. particularly with [BB12, KC08, ACD06].

5.1 Motivation

Observables of classical physical systems are typically understood to be functions on a suitable state-space of the system. This, however, presupposes that measurements on this systems do not influence the system in a significant way. Once this presupposition is abandoned, i.e. once measurements are understood as a process which may influence the system under investigation, untypical phenomena may occur even in classical systems.

E.g., we have seen in chapter 2 that the observables of NNs (which are completely classical in nature) can be non-commutative, a claim which is also supported through simulations of NNs in [AF06].

More general work in this direction has been carried out, e.g. in [bGFA12, bGA06, Fil12], where behavior of classical systems has been found which is similar to phenomena known from quantum-theory. This motivates the question of whether classical systems, as e.g. NNs, can violate Bell-type inequalities, which is to be answered in this chapter.

A second motivation of why an inquiry into Bell-type inequalities on NNs is legitimate relates to cognitive sciences, particularly to psychology. Since several years, experimental data have been obtained in psychology which suggests to model cognitive processes with a mathematical formalism akin to the mathematics used in quantum theory [BB12, AR12, AA95, BWT01, AFR04, DLM08, ABF$^+$08, BKNM09, PB09, BPFT11], a fact which even starts to be noticed by a broader science community [Mus12].

Part of this field is devoted to the study of violations of Bell-type inequalities, e.g. the CHSH-inequality, in psychological experiments [BB12, KC08, ACD06]. An investigation of whether classical systems, as given e.g. by NNs, allow for a violation of Bell-type inequalities may contribute to understand of those findings, particularly since cognitive processes are sometimes modeled with NNs (c.f. connectionism [Gar10] and [KSJ00]).

Thus, in this chapter, we consider Bell-type inequalities on NNs. In the following section, an introduction to the CHSH-inequality is given.

5.2 Introduction

In this chapter, we show that NNs in experimental situations may violate the CHSH-inequality. From a logical point of view, the CHSH-inequality follows from two major assumptions, as will be obvious in Sec. 5.3. They will be introduced in the following paragraphs. Before doing so, we emphasize that the CHSH-inequality does not refer to one individual measurement. Rather, it is a statistical inequality which can only be tested on an ensemble of systems or (equivalently) on individual systems which can be prepared in the same state an arbitrary number of times.

The first of the two assumptions which allow to derive the CHSH-inequality is (in a physical context) called **realism**. It assumes that there exists a parameter λ which determines the out-

come of measurements,[1] and that the probability distribution of λ over the ensemble of systems used to test the CHSH-inequality, ρ, is independent of parameters which can be chosen in a measurement, i.e. $\rho = \rho(\lambda)$.

The second of the two assumptions is referred to as 'locality'. Since the word 'locality' may easily provoke terminological misunderstanding, we will use a different term in the following. We say that two measurements are **independent** if there exist no means of mutual influence, i.e., if one measurement cannot in any way influence the other and vice versa. E.g., special and general relativity allow for *independent* measurements to be carried out: Two measurements are *independent* if their distance is space-like. However, if two measurements are not separated by a space-like distance, they are *not* independent: there exist means of mutual influence.

Thus, in classical physics, both independent and not independent measurement-pairs exist. Given realism as defined above, the CHSH-inequality has to hold for measurements which are *independent*. This claim will be proven in Sec. 5.3. I.e., from a logical point of view

$$\Big[\text{Measurement-independence \& Realism}\Big] \Rightarrow \text{CHSH-inequality holds.} \tag{5.1}$$

A violation of the CHSH-inequality therefore implies that at least one of the two assumptions is wrong. Since experiments have been carried out which violate the CHSH-inequality, we know that whichever theory correctly describes the part of reality which is probed by those experiments cannot obey both assumptions.

In the above-mentioned psychological experiments, measurements, which are performed in order to evaluate whether the CHSH-inequality, or similar Bell-type inequalities, are violated, are not separated by a space-like distance. I.e., even classical physical theories, as e.g. special relativity, do *not* predict the measurements to be *independent*. Therefore, there is no reason to expect the CHSH-inequality to hold for those measurement.

Since NNs might underlie cognitive processes (as conjectured, e.g., from both a connectionist [Gar10] and neuroscientific stance) this chapter focuses on the CHSH-inequality in connection with measurements on NNs. The measurements are a) understood - as throughout this thesis - to consist of showing an input to the NN under investigation and registering the output and b) not assumed to be separated by a space-like distance. Thus, they are conceptually equivalent to the measurements as relevant in the above-mentioned field.

We emphasize that this chapter does not aim to contribute anything to the quantum physics community. Classical (i.e, non-quantum) physics is assumed in all considerations.

5.3 Derivation of the CHSH inequality

In this section, we offer a derivation of the CHSH-inequality. It has been developed by Clauser, Horne, Shimony and Holt in 1969 [CHSH69]. We follow the derivation given in the original publication [CHSH69] since it is the conceptually clearest derivation known to the author. There

[1] In the original publication of Clauser, Horne, Shimony and Holt, λ denoted the hidden variables [CHSH69].

are several derivations which are easier to follow, e.g. the ones offered in [Per95, NC00, d'E79]. However, this increase in simplicity is traded against some loss of conceptual clarity, as exemplified by the appearance of counterfactual equations.[2]

Instead of treating "correlated pairs of particles" [CHSH69], we will consider a system S comprising at least two subsystems on which measurements can be performed.

Suppose that either an ensemble of systems S as specified in [CHSH69] is given or, equivalently, a single system S is available for repetitive inquiry which can be brought into its initial state at any time. Two measurements are carried out on subsystems of S, denoted M_a^A and M_b^B respectively, where a and b are parameters of the measurements which can be chosen freely. The result of each measurement, denoted by $A(a)$ and $B(b)$, respectively, is either $+1$ or -1. Suppose that the outcome of the two measurements is determined by a finitely large set of parameters, which we denote by λ.

As noted on p. 79, in order to derive the CHSH-inequality, we assume the two measurements to be **independent**, i.e. we assume that there is no possible interaction between M_a^A and M_b^B. This assumption implies that the result of M_a^A *cannot* depend on the choice of b and the result of M_b^B *cannot* depend on the choice of a, thus

$$A(a) = A(a, \lambda) \quad \text{and} \quad B(b) = B(b, \lambda), \tag{5.2}$$

instead of, e.g., $A(a) = A(a, b, \lambda)$.

Furthermore, we assume that the set of parameters λ (the 'state' of the system) is not depending on the choice of a or b. This implies that the (normalized) probability distribution ρ, which describes the ensemble of systems, only depends on λ:

$$\rho = \rho(\lambda). \tag{5.3}$$

As mentioned in Sec. 5.2, for a physical system, as e.g. a correlated pair of particles, this property is called (Einstein-) **realism**.

Let Γ be the space on which λ is defined. We define the correlation function between measurement-results of the two measurements as

$$P(a, b) = \int_\Gamma A(a, \lambda) B(b, \lambda) \rho(\lambda) d\lambda. \tag{5.4}$$

Using the fact that $A(a, \lambda), B(b, \lambda) \in \{+1, -1\}$, thus that $\frac{1}{B(b,\lambda)} = B(b, \lambda)$ and $|A(a, \lambda)B(b, \lambda)| = 1$, as well as the normalization of $\rho(\lambda)$, $\int_\Gamma \rho(\lambda) d\lambda = 1$, we obtain

$$\begin{aligned} |P(a, b) - P(a, c)| &\leq \int_\Gamma |A(a, \lambda)B(b, \lambda) - A(a, \lambda)B(c, \lambda)| \rho(\lambda) d\lambda = \\ &= \int_\Gamma |A(a, \lambda)B(b, \lambda)| [1 - B(b, \lambda)B(c, \lambda)] \rho(\lambda) d\lambda = \\ &= \int_\Gamma [1 - B(b, \lambda)B(c, \lambda)] \rho(\lambda) d\lambda = 1 - \int_\Gamma B(b, \lambda)B(c, \lambda) \rho(\lambda) d\lambda. \end{aligned} \tag{5.5}$$

[2] Counterfactual equations are given, e.g., if the results of two counterfactual measurements (i.e. of two mutually exclusive measurements) are contained in one equation.

Since $P(b,b') \in [-1,+1]$, define δ such that

$$P(b,b') = 1 - \delta. \tag{5.6}$$

Let $\Gamma_\pm$ be regions of Γ defined by $\Gamma_\pm = \{\lambda | A(b',\lambda) = \pm B(b,\lambda)\}$, where b and b' are parameters of the measurements M^A and M^B, respectively. Using $\int_\Gamma .. = \int_{\Gamma_+} .. + \int_{\Gamma_-} ..$ together with the normalization of ρ and the definition of $\Gamma_\pm$, we find

$$\begin{aligned}
\delta &= 1 - P(b',b) = 1 - \int_\Gamma A(b',\lambda)B(b,\lambda)\rho(\lambda)d\lambda = \\
&= \int_\Gamma \rho(\lambda)d\lambda - \int_{\Gamma_+} A(b',\lambda)B(b,\lambda)\rho(\lambda)d\lambda - \int_{\Gamma_-} A(b',\lambda)B(b,\lambda)\rho(\lambda)d\lambda = \\
&= \int_{\Gamma_+} \rho(\lambda)d\lambda + \int_{\Gamma_-} \rho(\lambda)d\lambda - \int_{\Gamma_+} (+1)\rho(\lambda)d\lambda - \int_{\Gamma_-} (-1)\rho(\lambda)d\lambda = \\
&= 2\int_{\Gamma_-} \rho(\lambda)d\lambda.
\end{aligned} \tag{5.7}$$

This yields

$$\begin{aligned}
&\int_\Gamma B(b,\lambda)B(c,\lambda)\rho(\lambda)d\lambda = \\
&= \int_{\Gamma_+} B(b,\lambda)B(c,\lambda)\rho(\lambda)d\lambda + \int_{\Gamma_-} B(b,\lambda)B(c,\lambda)\rho(\lambda)d\lambda = \\
&= \int_{\Gamma_+} A(b',\lambda)B(c,\lambda)\rho(\lambda)d\lambda + \int_{\Gamma_-} (-A(b',\lambda))B(c,\lambda)\rho(\lambda)d\lambda = \\
&= \int_{\Gamma_+} A(b',\lambda)B(c,\lambda)\rho(\lambda)d\lambda + \int_{\Gamma_-} (-A(b',\lambda))B(c,\lambda)\rho(\lambda)d\lambda + \\
&\qquad + \int_{\Gamma_-} A(b',\lambda)B(c,\lambda)\rho(\lambda)d\lambda - \int_{\Gamma_-} A(b',\lambda)B(c,\lambda)\rho(\lambda)d\lambda = \\
&= \int_\Gamma A(b',\lambda)B(c,\lambda)\rho(\lambda)d\lambda - 2\int_{\Gamma_-} A(b',\lambda)B(c,\lambda)\rho(\lambda)d\lambda \geq \\
&\geq P(b',c) - 2\int_{\Gamma_-} |A(b',\lambda)B(c,\lambda)|\rho(\lambda)d\lambda = \\
&= P(b',c) - 2\int_{\Gamma_-} \rho(\lambda)d\lambda = P(b',c) - \delta.
\end{aligned} \tag{5.8}$$

Together with (5.6) and (5.5), this gives the **CHSH-inequality**

$$|P(a,b) - P(a,c)| \leq 2 - P(b',b) - P(b',c). \tag{5.9}$$

Relabelling the variables and reordering terms gives

$$|P(a,b) - P(a,b')| + P(a',b) + P(a',b') \leq 2. \tag{5.10}$$

5.4 CHSH and arbitrary systems

Sec. 5.3 has made it obvious that the CHSH-inequality is a consequence of the following two assumptions:

(I) **Measurement-independence** of the relevant measurements: The choices of parameters a and b do *not* influence the respective other measurement, i.e. $A = A(a, \lambda)$, $B = B(b, \lambda)$.

(II) **Realism**: The probability-distribution describing the system(s) under investigation, ρ, is depending on λ only, and not on a or b.

Thus, if measurements on an *arbitrary* system S violate the CHSH-inequality (5.10), this implies that at least one of the assumptions (I) or (II) must be rejected.

5.5 CHSH and NNs

Suppose the system S specified above is a NN. It has already been noted that - in accordance with [BB12, KC08, ACD06] - the measurements on NNs are not assumed to be *independent* in a physical sense - i.e., they are not separated by a space-like distance. We presuppose the following scenario:

> The measurements are performed **coincidently**. This means that the input-neurons of the NN are separated into two sets, one belonging to measurement M_a^A and the other to measurement M_b^B. The parameters a and b are thus associated with different inputs which can be shown to the NN. The result of a measurement is the output the NN displays, where again one part of the output-neurons is associated with M^A and the other part with M^B. Since the CHSH-inequality has been derived for measurement-results being $+1$ or -1, we have to associate the outputs of the NN with either of the two. Thus, a scenario where coincident measurements are performed is of the following kind:
>
> – At time t_1 input a is shown to the input-neurons associated with M_a^A, and input b is shown to the input-neurons associated with M_b^B.
>
> – At time t_2 with $t_2 > t_1$, the result of the measurements is evaluated, i.e., the output of the NN is observed.

In accordance with Sec. 5.3, let $A(a, \lambda)$ and $B(b, \lambda)$ refer to the results of measurements M_a^A and M_b^B.

For the following reasons we cannot expect assumptions (I) or (II) to hold, i.e., we cannot expect the CHSH-inequality - or any Bell-type inequality - to be obeyed.

(I) The underlying structure of a NN is a graph (c.f. Ch. 1). If this graph is connected, as generally the case, information given to the NN (input) may spread globally. Thus, either of the inputs a and b may have an influence on the output-neurons associated with the respective other measurement. Thus, the measurements M_a^A and M_b^B are in general *not independent*.

(II) As we have stressed throughout this work, measurements performed on NNs in experimental situations are invasive, i.e., they change the state of the system (c.f. p. 1 or Sec. 1.2). Thus, we *cannot* assume that the probability distribution of the parameters λ, as given by ρ, is independent of the choice of inputs a and b, i.e., realism is violated.

The following theorem proves this point.

Theorem 3

In general, NNs can violate the CHSH-inequality.

Proof We construct a NN which violates he CHSH-inequality.

Consider a NN comprising four neurons $v_1, ..., v_4$, as illustrated in Fig. 5.1. Let M_a^A be a measurement presenting input on v_1 and registering the output of v_3. Let M_b^B be a measurement presenting input on v_2 and registering the output of v_4. Assume the inputs are presented at t_1 and the output is evaluated at $t_2 := t_1 + 2$.

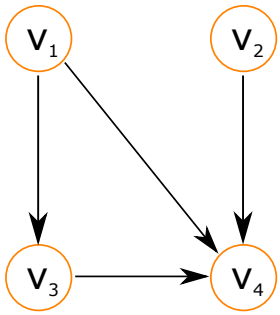


Fig. 5.1: Illustration of the NN used in the proof of theorem 3.

Let the neuron-states be $+1$ or -1, i.e. $I = \{+1, -1\}$ and $a, b \in \{+1, -1\}$. The following equations specify the state of each neurons at time $t + 1$ depending on the states of those neurons, which are connected to it, at time t. (I.e., the following equations specify the update-rule (p. 7) of the individual neurons.)

For every time t, let

$$\begin{aligned} u_{t+1}(v_1) &:= u_t(v_1), \\ u_{t+1}(v_2) &:= u_t(v_2), \\ u_{t+1}(v_3) &:= u_t(v_1), \\ u_{t+1}(v_4) &:= \mathrm{sgn}(u_t(v_1) + u_t(v_2)) \cdot u_t(v_3), \end{aligned}$$

where $\mathrm{sgn}(x)$ gives the sign of x with $\mathrm{sgn}(0) := +1$. The dynamics of this NN are illustrated in table 5.1.

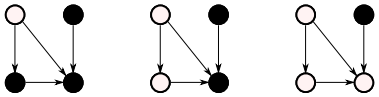
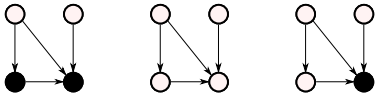
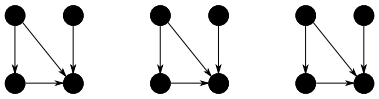
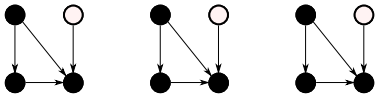
Parameters chosen for the measurements M_a^A and M_b^B respectively	t_1	t_1+1	$t_2=t_1+2$	$P(a,b)$
(-1, +1)				+1
(-1, -1)				-1
(+1, +1)				+1
(+1, -1)				+1

Table 5.1: Illustration of the dynamics of the NN used in the proof of theorem 3. Filled circles represent a neuron-state of $+1$, empty circles represent a neuron-state of -1. All neurons are assumed to occupy the state $+1$ before the onset of measurements, but this choice does not influence the final configuration at t_2.

Thus for the choices

$$a = -1,\ a' = +1$$
$$b = +1,\ b' = -1,$$

this NN violates the CHSH-inequality (5.10) maximally (i.e., even beyond the so-called quantum bound of $2 \cdot \sqrt{2}$):

$$|P(a,b) - P(a,b')| + P(a',b) + P(a',b') = 4.$$

Q.e.d.

Note that measurements can also be performed consecutively. It is straightforward to show that consecutive measurements on NNs may also violate the CHSH-inequality. However, in such a case, the temporal Bell inequalities are of interest rather than the CHSH-inequality. Further research in this direction can be found in [AF10] and [AF11].

Summary 10: NNs and the CHSH inequality

In this chapter, we elaborate on the relevance of the CHSH inequality w.r.t NNs. A motivation to do so can be found at the beginning of the chapter, p. 77.

We find that measurements, which are performed on an individual NN, can violate the CHSH-inequality. An explanation thereof can be found on p. 82. The reason lies mainly in

the fact that such measurements are not independent (this term is defined on p. 79): there exist means of mutual influence between two measurements.

In order to prove this, we construct a NN which violates the CHSH-inequality (maximally). It turns out that already very small NNs, consisting, e.g., of four nodes, suffice to do so (c.f. Fig. 5.1).

Conclusion and Outlook

In this thesis we define a notion of measurement on Neural Networks (NNs) which differs from what is conventionally understood as a 'measurement' on classical systems. Namely, a measurement consists of both, presentation of an input and registration of an output, where the latter is the 'result' of the measurement.

Thus, this notion of measurement is a process rather than an "observation". The investigations into the structure of such measurements, into the properties of observables associated with those measurements and into the relation between measurements and states all contribute to the question of which consequences a - compared to conventional classical systems - more general notion of measurement has.

First of all, the question exists of whether algebraic observables can be constructed for arbitrary systems at all. We show that for NNs and measurements as defined, this is indeed possible, which opens the way to study the properties of those observables.

The latter is of interest from many perspectives. First, it is an open question of whether the observables of all physical systems form a C^*-algebra. This is certainly so for quantum theory and classical physics. We find that the observables of a NN do not form a C^*-algebra, independent of the definitions of any binary relations on the set of observables. Thus, this thesis suggests that the answer to the above question is 'no': There are systems whose observables do not form a C^*-algebra.

Second, this generates the question of which structure can be assumed to hold for arbitrary systems. Or put differently, which axioms describe any system, on which investigations can be performed, correctly. Research in this direction is being carried out, and several groups offer answers to this question. We chose one promising approach, called 'Generalized Quantum Theory' and evaluate first whether a treatment as suggested by this approach can be carried out on NNs, and second whether the proposed axioms hold on NN. Whereas we find that the answer to the first question is 'yes', some of the proposed axioms do not hold for NNs.

Thus, this thesis contributes an extensive example to the question of which properties observables of physical and non-physical systems can satisfy, which is not yet covered by an axiomatic scheme.

In the second part of this thesis, we study whether the notion of states, and the possibility to perform inquiries into which state a NN occupies, is effected by the definition of measurement

as well. It turns out that this is the case. Whereas conventional notions of measurements are compatible with a notion of state which is based on the states of the individual neurons (e.g., probability distributions over the set of neurons), we find that the definition of measurement, which is presupposed here, suggests that attractors of the NN constitute its states. This is so because a) individual states of all neurons cannot be identified by measurements, b) a description of the behavior of the NN in terms of attractors is feasible and c) it is possible to identify the attractor, which a NN occupies, through series of consecutive measurements.

It is the latter point which constitutes most of the second part of this thesis. Explicit methods are constructed which allow to carry out this identification. In a more general perspective, it is a contribution to the question of how a state of a system can be identified by histories of measurements.

We show that the methods, which are proposed to identify an attractor, offer a novel way of studying NNs in experimental situations. However, those methods are proven to work only for NNs which are compatible with the definitions of this thesis. Even though the latter have carefully been chosen to be as general as possible, four restrictions had to be added to make a treatment as outlined above feasible. It is unlikely, e.g., that biological NNs comply with those restrictions.

This constitutes the most important outlook with respect to further research along the lines of this thesis. If some of the four constraints, e.g., the finiteness of the number of states, which individual neurons may occupy, or the necessity of a deterministic update-rule, can be released, the methods of this thesis might offer interesting insights into NNs even in biology or neuroscience. This might be the case, e.g., if it can be assured in some way that the output of a NN - not the NN itself - behaves as is the case in this thesis, i.e., most importantly, that the attractor, which a NN occupies (including strange attractors or limit tori) can be associated with the output of a NN (including series of "direct" output) through a *function*. Furthermore, measurement-errors or noise would have to be included into the concepts.

Concerning the conceptual investigations of this thesis, further research could be carried out in two directions. On the one hand, some restrictions of the definitions could be dropped in order to evaluate whether the properties of observables change and which axioms still hold for the system. On the other hand, more restrictions could be added in order to investigate the full range of unexpected behavior of NNs.

Finally, one outlook of this thesis concerns results which have not been mentioned so far in this section: A proof that NNs (given measurements as above) can violate the CHSH-inequality, which is of interest for a particular field in psychology. Future work could evaluate in how far other quantum-like behavior, which is found in this field, can be explained on the basis of NNs, or, more importantly, what such (empirical) findings imply with respect to the system which underlies the relevant cognitive processes.

Bibliography

[AA95] Diedrik Aerts and Sven Aerts. Applications of quantum statistics in psychological studies of decision processes. *Foundations of Science*, 1(1):85–97, 1995.

[ABF+08] Harald Atmanspacher, Michael Bach, Thomas Filk, Jürgen Kornmeier, and Hartmann Römer. Cognitive time scales in a necker-zeno model for bistable perception. *Open Cybernetics and Systemics*, 2:234–251, 2008.

[ACD06] Diederik Aerts, Marek Czachor, and Bart D'Hooghe. Towards a quantum evolutionary scheme: Violating bell's inequalities in language. In Nathalie Gontier, Jean Paul Bendegem, and Diederik Aerts, editors, *Evolutionary Epistemology, Language and Culture*, volume 39 of *Theory and Decision Library A*, pages 453–478. Springer Netherlands, 2006.

[AF06] Harald Atmanspacher and Thomas Filk. Complexity and non-commutativity of learning operations on graphs. *Biosystems*, 85(1):84–93, 2006.

[AF10] H. Atmanspacher and T. Filk. A proposed test of temporal nonlocality in bistable perception. *Journal of Mathematical Psychology*, 54:314–321, 2010.

[AF11] Harald Atmanspacher and Thomas Filk. Options for testing temporal bell inequalities. In P. Bruza et al., editor, *Quantum Interaction*, pages 128–137. Springer, 2011.

[AFR04] Harald Atmanspacher, Thomas Filk, and Hartmann Römer. Quantum zeno features of bistable perception. *Biological Cybernetics*, 90:33–40, 2004.

[AFR06] Harald Atmanspacher, Thomas Filk, and Hartmann Römer. Weak quantum theory: Formal framework and selected applications. In G. Adenier, Y. Khrennikov, and T. M. Nieuwenhuizen, editors, *Quantum Theory: Reconsideration of Foundations*. American Institute of Physics, 3 edition, 2006.

[Ami89] Daniel J. Amit. *Modeling Brain Function*. Cambridge University Press, 1989.

[AP05] Harald Atmanspacher and Hans Primas. Epistemic and ontic quantum realities. In Andrei Khrennikov, editor, *Foundations of Probability and Physics*. AIP Press, 2005.

[AR12] H. Atmanspacher and H. Römer. Order effects in sequential measurements of non-commuting psychological observables. *Journal of Mathematical Psychology*, 56(4):274–280, 2012.

[ARW02] Harald Atmanspacher, Hartmann Römer, and Harald Walach. Weak quantum theory: Complementarity and entanglement in physics and beyond. *Foundations of Physics*, 32:379–406, 2002.

[BB12] Jerome R. Busemeyer and Peter D. Bruza. *Quantum Models of Cognition and Decision*. Cambridge University Press, 2012.

[BEH99] Jiri Blank, Pavel Exner, and Miloslav Havlicek. *Hilbert Space Operators in Quantum Physics*. AIP Press, 1999.

[bGA06] Peter beim Graben and Harald Atmanspacher. Complementarity in classical dynamical systems. *Foundations of Physics*, 36:291–306, 2006.

[bGFA12] Peter beim Graben, Thomas Filk, and Harald Atmanspacher. Epistemic entanglement due to non-generating partitions of classical dynamical systems. *International Journal of Theoretical Physics, in press*, 2012.

[BKNM09] Peter Bruza, Kirsty Kitto, Douglas Nelson, and Cathy McEvoy. Is there something quantum-like about the human mental lexicon? *Journal of Mathematical Psychology*, 53(5):362–377, 2009.

[BPFT11] Jerome R. Busemeyer, Emmanuel Pothos, Riccardo Franco, and Jennifer S. Trueblood. A quantum theoretical expanation for probability judgment errors. *Psychological Review*, 108:193–218, 2011.

[BS03] Stefan Bornholdt and Heinz Georg Schuster. *Handbook of Graphs and Networks*. Wiley-VCH, 2003.

[BSMM05] Ilja N. Bronstein, Konstantin A. Semendjajew, Gerhard Musiol, and Heiner Mühlig. *Taschenbuch der Mathematik*, volume 6. Harri Deutsch, 2005.

[BWT01] Jerome R. Busemeyer, Zheng Wang, and James T. Townsend. Quantum dynamics of human decision making. *Journal of Mathematical Psychology*, 50:220–241, 2001.

[By03] Yaneer Bar-yam. *Dynamics Of Complex Systems*. Westview Press, 2003.

[CHSH69] John F. Clauser, Michael A. Horne, Abner Shimony, and Richard A. Holt. Proposed experiment to test local hidden-variable theories. *Phys. Rev. Lett.*, 23:880–884, 1969.

[DA01] Peter Dayan and Larry F. Abbott. *Theoretical Neuroscience*. MIT Press, 2001.

[d'E79] Bernard d'Espagnat. The quantum theory and reality. *Scientific American*, 1979.

[DLM08] Vladimir I. Danilov and Ariane Lambert-Mogiliansky. Measurable systems and behavioral sciences. *Mathematical Social Sciences*, 55:315–340, 2008.

[Fil05] Thomas Filk. Grundlagen und Probleme der Quantenmechanik. *Lecture notes: `http://omnibus.uni-freiburg.de/~filk/Skripte/Texte/Quanten.pdf`*, 2005.

[Fil12] Thomas Filk. Quantum-like behavior of classical systems. In J.Busemeyer et al., editor, *Proceedings of the Conference QI2012, Paris*. in press., 2012.

[Fis05] Gerd Fischer. *Lineare Algebra*. vieweg, 15 edition, 2005.

[FR11] Thomas Filk and Hartmann Römer. Generalized quantum theory: Overview and latest developments. *Axiomathes*, 21:211–220, 2011.

[Gar98] G. David Garson. *Neural Networks: An Introductory Guide for Social Scientists*. Sage Publications Ltd, 1998.

[Gar10] James Garson. Connectionism. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Winter 2010 edition, 2010.

[GK02] Wulfram Gerstner and Werner Kistler. *Spiking Neuron Models*. Cambridge University Press, 2002.

[GN43] Izrail Moiseevich Gel'fand and Mark Aronovich Naimark. On the imbedding of normed rings into the ring of operators in Hilbert space. *Rec. Math. [Mat. Sbornik] N.S.*, 1943.

[GS09] Thilo Gross and Hiroki Sayama, editors. *Adaptive Networks: Theory, Models and Applications*. Springer, 2009.

[Har11] Tero Harju. Lecture notes on graph theory. *Lecture notes: `http://users.utu.fi/harju/graphtheory/graphtheory.pdf`*, 1994-2011.

[Hay98] Simon Haykin. *Neural Networks: A Comprehensive Foundation*. Prentice Hall, 2nd edition, 1998.

[Hay08] Simon O. Haykin. *Neural Networks and Learning Machines*. Prentice Hall, 3rd edition, 2008.

[HS85] Friedrich Hirzebruch and Winfried Scharlau. *Einführung in die Funktionalanalysis*. B.I.-Hochschuhltaschenbücher, 1985.

[KC08] Andrei Khrennikov and Elio Conte. A preliminary experimental verification on the possibility of Bell inequality violation in mental states. *Neuroquantology*, 6(3):214–221, 2008.

[Kni07] Simeon Knieling. *Einführung in die Modellierung künstlich neuronaler Netzwerke*. WiKu-Verlag Verlag für Wissenschaft und Kultur, 2007.

[KSJ00] Eric Kandel, James Schwartz, and Thomas Jessell. *Principles of Neural Science*. McGraw-Hill, 2000.

[Lan98] N. P. Landsman. Lecture notes on C*-algebras, Hilbert C*-modules, and quantum mechanics. *ArXiv Mathematical Physics e-prints*, July 1998.

[Mat98] Martin Mathieu. *Funktionalanalysis: Ein Arbeitsbuch*. Spektrum Akademischer Verlag, 1998.

[MP43] Warren McCulloch and Walter Pitts. A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biology*, 5:115–133, 1943.

[Mus12] George Musser. Humans think like quantum particles. *Scientific American*, 2012.

[NC00] Michael A. Nielsen and Isaac L. Chuang. *Quantum Computation and Quantum Information*. Cambridge University Press, 2000.

[PB09] Emmanuel Pothos and Jerome R. Busemeyer. A quantum probability model explanation for violations of rational decision theory. *Proceesings of the Royal Society B*, 276(1165):2171–2178, 2009.

[Per95] Asher Peres. *Quantum Theory: Concepts and Methods*. Kluwer Academic Publishers, 1995.

[Pri90] Hans Primas. Mathematical and philosophical questions in the theory of open and macroscopic quantum systems. In A. I. Miller, editor, *Sixty-Two Years of Uncertainty: Historical, Philosophical, and Physical Inquiries into the Foundations of Quantum Mechanics*, New York, 1990. Plenum Press.

[RHB06] K. F. Riley, M. P. Hobson, and S. J. Bence. *Mathematical Methods for Physics and Engineering*. Cambridge University Press, 2006.

[RZ94] Heinz Rehkugler and Hans Georg Zimmermann. *Neuronale Netze in der Ökonomie. Grundlagen und finanzwirtschaftliche Anwendungen.* Vahlen, 1994.

[Sch98] Walter J. Schempp. *Magnetic Resonance Imaging*. Wiley-Liss, 1998.

[SK11] Ian H. Stevenson and Konrad P. Kording. How advances in neural recording affect data analysis. *Nat Neurosci*, 14(2):139–142, 2011.

[vHS94] J. Leo van Hemmen and Klaus Schulten. *Models of Neural Networks: Temporal Aspects of Coding and Information Processing in Biological Systems*. Springer, 1994.

[vN96] John von Neumann. *Mathematische Grundlagen der Quantenmechanik*. Springer, 1932 (reprint 1996).

[Wil96] Robin J. Wilson. *Introduction to Graph Theory*. Prentice Hall, 4th edition, 1996.

List of Figures

List of Tables

Erklärung

Ich versichere, dass ich diese Arbeit selbstständig verfasst habe und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe.

Heidelberg, den 25.10.2012 .....