Utrecht University

MASTER THESIS

HISTORY AND PHILOSOPHY OF SCIENCE

The Ontological Status of Information in Physics

Author
Nick WIGGERSHAUS
(5724767)

Supervisor
Dr. Guido
BACCIAGALUPPI

Second Reader
Prof. dr. Dennis
DIEKS

November 2017

Abstract

While the inclusion of information theoretical concepts into (quantum) physics has shown enormous success in recent years, the ontology of information remains puzzling. Therefore, this thesis aims to contribute to the debate about the ontological status of information in physics. Most of the recent debates have focused on syntactic information measures and especially Shannon Information, a concept originally stemming from Communication Theory. This thesis incorporates another syntactic information measure, the so far largely underrepresented notion of Algorithmic Information or Kolmogorov Complexity, a concept often applied in Computer Science. Shannon Information and Kolmogorov Complexity are linked through Coding Theory and have similar characteristics. Through the comparison of Shannon Information and Kolmogorov Complexity a framework is developed which analyses the respective information measures in relation to uncertainty and semantic information. In addition, this framework investigates whether information can be regarded a material entity and examines to what extent information is conventional. It turns out that in the classical case Shannon Information and Kolmogorov Complexity are both abstract and highly conventional entities, which must not be confused with uncertainty and do not bear any relations with semantic information. Virtually the same results are obtained in the quantum case, save for the high degree of conventionality; it is argued that Quantum Theory constrains the conventional choices of those who wish to use either theory.

Acknowledgements

First and foremost, I would like to express my gratitude to my supervisor Guido Bacciagaluppi. Without his support and the countless hours of fruitful discussions, I would not have been able to write this thesis. Unsuspecting I chose a topic with far reaching implications and connections to unknown research areas I did not anticipate in the beginning. Thank you for your patience and always having an open door!

Second, I would like to thank the various scholars who helped me–in one way or another–during my writing process. At an early stage, my second reader Dennis Dieks provided me with useful clarifications about Shannon Information and encouraged me to reach out to other scholars. In this context, I have to thank Olimpia Lombardi, for her kind responses and sending me the manuscripts of the international workshop: *What is Quantum Information?* (Buenos Aires. 2015); Christopher Timpson for sharing his insights about the topic through e-mail and in between conference talks in Oxford (March 2017); and Paul Vitany for clarifying important details about Kolmogorov Complexity through e-mail correspondence.

Third, I would like to thank my fellow students and people affiliated with the HPS Masters in Utrecht who I shared the experience of writing a thesis with and gave kind words of encouragement. Special thanks to Noelia Iranzo Ribera for her helpful comments and remarks regarding some of the chapters found this thesis.

Last, I am grateful for the enormous mental support of my parents, grandparents, friends, flatmates and coworkers. Thanks for your understanding and thoughtfulness when the mysteries surrounding the ontological status of information caused me hours of despair. *Vielen Dank!*

Contents

List of Figures

Chapter 1

Introduction

> "It is said that we live in an Age of Information, but it is an open scandal that there is no theory, nor even definition, of information that is both broad and precise enough to make such an assertion meaningful." [Goguen, 1997, p. 27]
>
> – *Goguen*

WHILE attending the conference of the *European Society of History of Science* (ESHS) in September 2016 in Prague and pondering about a suitable topic for this master thesis, the initial impulse came *Frans van Lunteren's* talk, professor of History of the natural sciences in Leiden. Van Lunteren suggested to compose a history of science textbook for science students, following the narrative that different technological advancements had enormous influence on scientific practice; think for instance how the steam engine influenced thermodynamics or how the universe is conceived as a 'giant clock' [van Lunteren, 2016].

Since the middle of the last century until today–so the claim–we are largely influenced by information technology and especially the role of the computer. And indeed, in our daily lives we are permanently surrounded by the term 'information'. What's more, many academic disciplines are in one way or another affected by some of the multitude concepts of 'information', too. One of these academic disciplines is physics, where the new areas of *Quantum Information Theory* (QIT) and *Quantum Information Science* (QIS) offer one of the most exciting and rapidly growing areas of

research in the field. Quite generally speaking, this thesis is concerned with information in physics.

This chapter aims to illuminate the wide ambiguity around the term, arguing that 'information' appears to refer to at least two quite distinct concepts. Beginning with a quick glance in the online Oxford dictionary[1] provides us with two main entries, according to which information is a mass noun defined as "[f]acts provided or learned about something or someone", and "[w]hat is conveyed or represented by a particular arrangement or sequence of things." While the first entry refers to a subjective-epistemic notion, the second entails the syntactic-quantitative and rather 'objective' concepts.[2] In order for the main argument of this paper to hold from the start, we have to identify that modern physics is concerned with *only* the latter of these concepts. However, often these two meanings of information are conflated; therefore, a short insight in the *begriffsgeschichte* of information is provided for clarification and subsequently concluded with the developments in the 20th century until today.

Embarking from there, we'll see that there are various notions of so called *syntactic information*. In this thesis we'll then pick out the arguably two most prominent syntactic information measures, *Shannon Information* and *Kolmgorov Compelxity*, and raise the question '*What's the ontological status of information in physics?*'. While Shannon Information and its 'quantum version' have already gained a lot of attention in the foundations of physics literature, only a few authors have philosophically examined Kolmogorov Complexity and its extensions to the quantum. On the one hand, contrasting both of these information measures yields a fruitful framework to examine Shannon Information from a so far unusual angle. On the other hand, the analysis of Kolmogorov Complexity allows this thesis to act as a starting point to close the 'gap of attention' between the two information measures when it comes to the ontological status of information in physics. Even though we begin to analyze each notion individually, we are going to point out conceptual

[1] URL= https://en.oxforddictionaries.com/definition/information (01/06/2017)
[2] 'Objective' in sense of intersubjective.

connections between Shannon Information and Kolmogorov Complexity, in order to shed light on the ontological status of information. We introduce the exact main conceptual framework for settling the main research question of this thesis at the end of the chapter. Before that, we start with a short briefing on ontology.

1.1 A few words about Ontology

Before beginning with the main part of thesis, it is advisable to leave a few words of clarification about how 'ontological status' ought to be understood here. Overall, ontology is the inquiry with the question of *what there is*. Ontologists want to give a catalog of the world's furniture and ask for instance, whether there exist numbers; Platonists may answer yes and Nominalists no. More generally, it is debated whether there exist abstract entities (such as numbers) or concrete entities.

However, for the main analysis of this thesis, we want to eschew most of these ontological debates and not digress from our actual topic by arguing over the pitfalls of various ontological positions. First and foremost, we are concerned with the metaontological question of categorizing the proposed notions of syntactic information in physics according to their ontological status. In other words, when examining the ontological status of information, we want to step back from any ontological position such as Platonism, Nominalism, or the question if ontology is solely a matter of convention, as choosing one of these metaphysical positions deserves an independent debate.

In order to meet the goal of this thesis–investigating the ontological status of information in physics–it is crucial to point out some of these basic ontological concepts and categories. This shall later allow us to analyze the various ontological claims about information in physics we are going to encounter and put them into perspective.

Abstract vs Concrete Entities

Let's start with the perhaps most basic and intuitive notion familiar from our experience–that of *concrete* objects, like apples, chairs or tigers. What all these things or objects have in common is that each of them is said to be non-repeatable (not multi-examplifiable) [MacLeod and Rubenstein, 2017], which means that they can't be spatio-temporally located in more than one place at a time. A single tiger, for instance, can't be at the zoo in Chicago and the Indian jungle at the same time; there can certainly be two *different* tigers at different locations at the same time though, one in Chicago and one in the tropical rainforest of Northeast India. Often these (material) objects or things are referred to as *individuals* or *particulars*,[3] which can be picked out from a certain class or category.

A diagnosis similar to the whereabouts of big cats and their categorization as tigers–or generally about not multi-examplifiable particulars and their resemblance–is one of the reasons why, historically, philosophers adhered to the notion of *universals*, a class of mind-independent entities, explaining the relations of qualitative identity among individuals [MacLeod and Rubenstein, 2017]. In contrast to particulars, universals are thought to be repeatable. Often universals account for the *properties* of objects, like the redness of apples or a tiger's property of being alive, striped and heavy. Holding the view that properties are in fact universals then allows us to conclude that the two tigers–the one in the zoo and the other in the jungle–are instances or *tokens* of the 'tiger'-*type* (i.e., if 'tiger' is recognized as a property itself, the property which characterizes all tigers). In philosophical terms we might say that a tiger (object) then *instantiates* various properties such as 'stripeness', 'heaviness', or the property of being a tiger.

However, denying that universals exist and believing that properties are merely particulars after all, is reflected in the so called *trope theory*. According to trope theory, the existence of universals is rejected and the world is held to consist (wholly or partly) of so called *tropes*, which are *abstract* particulars [Maurin, 2016]. A trope is a particular instance of a

[3]Examples for non-material-, i.e. abstract particulars are e.g., 'God' or 'souls'.

property, e.g., the specific redness of an apple.

Regardless whether one follows trope theory or adheres to the 'the classical distinction' of universals and particulars, the essential point for this thesis is that an object can evidently be spatio-temporal located, whereas its properties *can't* be localized like ordinary objects. So while we could certainly pet a tiger or eat an apple, we couldn't do the same with 'redness', 'aliveness'or the property that instantiates all tigers. Questions about location (in our case at least),[4] are thus only sensical for objects which *possess* properties and not the properties themselves; properties are *abstracta*.

1.2 Begriffsgeschichte

On the first page of his dissertation about the intellectual history and etymology of information [Capurro, 1978], Rafael Capurro warns us that:

> "The present everyday language term for information was initially used in a naive or rather nonreflective sense, e.g. as synonym for perception, knowledge, note, message, etc. At the same time, at the beginning of the thirties [of the 20th century], the word was annexed by telecommunications and by exclusion of all qualitative or respectively semantic aspects, which mark the everyday term of information, newly defined. The definition of the notion of information in telecommunications collided with the everday meaning. The suspicion of misguidance was confirmed, when the telecommunicational term of information was put in contact with semantic- and even pragmatic concerns.", (p.1).[5]

In fact, regarding the ancestry of the term information, we discover that only the epistemic notion has 'Latin roots and Greek origins', whereas the

[4] Some scholars might regard 'events' as yet another ontological category distinct from objects or properties. Although events are usually not regarded as concrete, they can be assigned with a spatio-temporal location.

[5] Own translation.

non-semantic notion is largely influenced by its emergence in the context of communication theory and computer science in the 20th century.

As Capurro points out, the term originates in the Latin $forma$, especially in the context of translations of philosophical works of Plato and Aristotle, covering a plurality of meanings such as $\varepsilon\iota\delta o\varsigma/eidos$ (essence), $\iota\delta\varepsilon\alpha/idea$ (idea), $\tau\upsilon\pi o\varsigma/typos$ (type), $\mu o\rho\varphi\eta/morphe$ (form) [Lyre, 1998], [Capurro and Hjorland, 2003], [Adriaans, 2013]. In Antiquity, the Latin *informatio* then had two fundamental meanings, i.e. i) "the action of giving a form to something material" and ii) "the act of communicating knowledge to another person" [Capurro, 2009]. In contrast to our contemporary understanding, the Latin term had thus an ontological and epistemological meaning, which nevertheless were closely related. In that regard, [Adriaans, 2013] e.g., points towards the important image of molding wax, used by various authors from Antiquity to the Early modern period. The fact that wax can be manipulated to have different shapes while keeping its volume and can be used to convey information–in sense of communicating knowledge–made it a rich analogy.

In the Middle Ages the term 'informatio' gained another meaning in pedagogical- and educational contexts. In the transition from the Middle Ages to the Early Modern Period and with the emergence of the various Latin-influenced European languages, the term was even further applied in different areas, such as in legal matters. However, with a more widespread use, under the influence of Empiricism and the decline of Scholasticism, 'informatio' slowly lost its ontological meaning (i.e. forming matter), retaining only its epistemic sense of communication and education [Capurro and Hjorland, 2003].

Developments in the 20th Century

Until the beginning of the 20th century, 'information' almost vanished completely from philosophical discourse. With a gradually shifted meaning towards views of knowledge in colloquial speech, information eventually became the abstract mass-noun we use today [Adriaans, 2013].[6]

[6]N.B. that 'information' isn't a mass-noun in every language. Think of Italian or German as a counter example.

In this form then, the term was chosen by 20th-century researchers of different scientific disciplines to adopt formal and 'objective' methods for measuring information.[7]

Historically, one aspect of syntactic information emerged in the context of communication theory in the Bell laboratories in the first half of the 20th century. Based on the work of Harry Nyquist (1889-1976) [Nyquist, 1924] and Ralph Hartley (1888-1970)[Hartley, 1928], Claude Shannon (1916-2001), published *A Mathematical Theory of Communication* [Shannon, 1948], in which he derived a probability based measure of information $H(X)$–called Shannon Information or Shannon Entropy–nowadays finding use in many scientific areas. Yet, at the time of the publication of his paper, allegedly Shannon himself struggled with the nomenclature of his newly found concept:

> "My greatest concern was what to call it. I thought of calling it 'information', but the word was overly used, so I decided to call it 'uncertainty'. When I discussed it with John von Neumann, he had a better idea. Von Neumann told me, 'You should call it entropy, for two reasons. In the first place your uncertainty function has been used in statistical mechanics under that name, so it already has a name. In the second place, and more important, nobody knows what entropy really is, so in a debate you will always have the advantage." [Tribus and McIrvine, 1971]

Whether the above protagonists indeed had such a conversation, is not our main concern. But the anecdote demonstrates important issues about the ambiguous terminology of $H(X)$ pointed out by Capurro above.

In the early 1960s, Ray Solomonoff (1926-2009) [Solomonoff, 1964], Andrey Kolmogorov (1903-1987) [Kolmogorov, 1965] and Gregory Chaitin (born 1947) [Chaitin, 1966], independently developed the notion of yet

[7] As an often cited example see for instance, statistican and biologists Ronald Fisher's (1890-1962) [Fisher, 1925] notion of *Fisher Information* in context of likelihood estimations. Because lack of space of this thesis, Fisher's measure of information will not be dealt with here any further. For some introductory reading though, see [Hilgevoord and Uffink, 1991].

another information measure, called *algorithmic information*. In the early 1960's, Solomonoff was the first to articulate the basic ideas of algorithmic information theory in the context of Artificial Intelligence (A.I.) research, trying to formulate a general theory of inductive reasoning.[8] However, at first Solomonoff's work regarding algorithmic information related to inductive reasoning went widely unnoticed. Motivated by Information Theory and problems of randomness, the approach by the great Soviet mathematician Kolmogorov to close a conceptual gap with Shannon's $H(X)$ in respect to single sequences, gained wider attention. For that reason, algorithmic information is today often referred to as *Kolmogorov Complexity.*[9]

Even though the notions of syntactic information were developed by different and seemingly independent motivations, Chaitin (one of the 'inventors/discoverers' of algorithmic information) remarks

> "Algorithmic information theory (AIT) is the result of putting Shannon's information theory and Turing's computability theory into a cocktail shaker and shaking vigorously."[10]

Unfortunately though, vigorously mixing various (ambiguously defined) concepts like uncertainty, entropy, probability, algorithmic, and semantically based 'information', doesn't help much in establishing conceptual clarification in science. As we are concerned with 'information in physics', it is time to look at what the physicists, the philosophers, and other scholars thereof have to say about the matter.

[8] As a historical side note, it is interesting to point out that in 1956 Solomonoff was one of the ten scholars of the Dartmouth Summer Study Group on A.I. (where, by the way, the term A.I. was coined), which was co-organized by C.E. Shannon (For more details see [Solomonoff, 1997] and [Vitanyi, 2010]).

[9] Only from 1968 onward did Kolmogorov start referring to Solomonoff's previous work, stating that he had not been aware of his developments earlier [Vitanyi, 2010]. Of course, many others like Leonid Levin, Peter Gacs, Gregory Chaitin, etc. made important contributions, too.

[10] Quote retrieved from [https://www.cs.auckland.ac.nz/research/groups/CDMTCS/docs/ait.php] (21/07/2017).

1.3 Information in Physics

While nowadays Information Theory has its own established standing, the usage of 'information' in physics is a more recent development and has arisen as a widely recognized scientific field since the early 1990s. Since the first developments of the computer in mid 20th century, the number of transistors in dense integrated circuits almost doubled every two years, an observation known as Moore's Law. While the dimensions of the components come ever closer to scales where quantum theory plays a crucial role, the limits of communication systems are being investigated by the effects of quantum mechanics on information transmission, too [Lloyd, 2009]. Quantum computation and quantum cryptography are only some of the new promising technological advances. Furthermore, many lines of quantum information theoretic research hope to overcome the puzzles of quantum mechanics and provide an entirely new foundation for the field.

However, increasingly successful application of information to various areas doesn't prevent conceptual confusion about the fundamental nature of information per se. As sketched out in the sections above, information is a highly ambivalent term and mingling it with no less ambiguous terms (like uncertainty, entropy, etc.) when brought in contact with physics hasn't proved helpful to untangle these conceptual confusions. Below, some of the most prominent claims about the ontological status of information in physics are presented.

Wheeler's *It from Bit*–Informational Immaterialism?

'It from bit'; probably one of the most notable slogans regarding information in physics, was coined by John Archibald Wheeler (1911-2008). In his own words, he claimed

> "It from bit. Otherwise put, every it–every particle, every field of force, even the spacetime continuum itself–derives its function, its meaning, its very existence entirely–even if in some contexts indirectly–from the apparatus-elicited answers

> to yes or no questions, binary choices, bits. It from bit symbolizes the idea that every item of the physical world has at bottom–at a very deep bottom, in most instances–an immaterial source and explanation; that what we call reality arises in the last analysis from the posing of yes-no questions and the registering of equipment-evoked responses; in short, that all things physical are information-theoretic in origin and this is a *participatory universe*." [Wheeler, 1989, p.5]

According to this position, information is the most fundamental, yet apparently immaterial, essence of the world. Information theoretic considerations justify experimentally posed yes or no questions which ultimately disclose the being of all material objects. In addition, forms of informational immaterialism seem to appeal to a certain form of epistemic uncertainty, as the posing of 'yes or no questions' is required.

While Wheeler's position of a participatory universe appears radically different from most mainstream views, he is not the only scholar proposing ideas in such a vein. Anton Zeilinger, one of the leading experimental physicists in the field of Quantum information, endorses a similar view (see e.g. [Zeilinger, 2004]). Even more recently, various essays in the Foundational Questions Institute (FQXi) essay competition [Aguirre et al., 2015], have argued in some way or another in favor of informational immaterialism (see e.g., Giacomo Mauro D'Ariano's *It from Quibit* (pp.25-35)). Today informational immaterialism still appears to remain a prominent position about the ontology of information in physics.

Landauer–Is Information Physical?

Another slogan we often encounter is 'Information is physical'. Such a position was most notably represented by Rolf Landauer (see [Landauer, 1991] and [Landauer, 1996]). Landauer comes to conclusions like:

> "Information is inevitably inscribed in a physical medium. It is not an abstract entity. It can be denoted by a hole in a punched card, by the orientation of a nuclear spin, or by the

> pulses transmitted by a neuron. The quaint notion that information has an existence independent of its physical manifestation is still seriously advocated. This concept, very likely, has its roots in the fact that we were aware of mental information long before we realized that it, too, utilized real physical degrees of freedom." [Landauer, 1999, p.64].

On this view, information is *not an abstract entity*, dependent on a *concrete* physical medium. One often finds similar notions embraced in engineer textbooks, where it is claimed that information is an entity that 'flows' in communication systems.

Timpson–Is Information Abstract?

In [Timpson, 2004], [Timpson, 2006], [Timpson, 2010], [Timpson, 2013] Christopher Timpson has argued for information being an abstract entity. As illustration, Timpson concludes that

> "[i]nformation$_t$, what is produced by a source, or what is transmitted, is not a concrete thing or a stuff. It is not so, because, as we have seen, what is produced/transmitted is a sequence type and types are abstracta. They are not themselves part of the contents of the material world, nor do they have a spatio-temporal location." [Timpson, 2006, p.27][11]

On such a view 'abstract' has to be understood as non-concrete (or non-material) as presented in the section about ontology above.

1.4 The Conceptual Framework of this Thesis

After our brief examination of the 'information-literature', let us now develop the framework of this thesis. Motivated by the sheer amount of

[11] Timpson's notion of information$_t$ can be understood as our notion of syntactic information.

the largely diverging interpretations of information in physics, we want to deepen our understanding about the ontological status of Shannon Information and Kolmogorov Complexity. In that regard it is helpful to recapture, as Tim Maudlin expressed, that:

> "Ontology is the most generic study of what exists. Evidence for what exists, at least in the physical world, is provided solely by empirical research. Hence the proper object of most metaphysics is the careful analysis of our best scientific theories (and especially of fundamental physical theories) with the goal of determining what they imply about the constitution of the physical world." [Maudlin, 2007, p.104]

In order to find an answer to our main research question, we have to analyze what information based theories in physics imply about the constitution of (syntactic) information.

However, as we have seen, there's large potential to confuse our examination of such theories before actually getting started, because the term is highly overworked and has at least two different meanings. Additionally, it's not always clear which information measure is referred to and what each formalism implies about the ontological status of information. For avoiding the pitfalls of ambiguous terminology, we have to be aware to be committed to the following two steps of analysis.

Step 1: Formal Introduction of Information Measures

At first, we have to clarify which kind of information measure we are referring to, i.e. what we are actually talking about. The previous sections gave us an idea that we ought to be careful about ambiguous terminology and a thus potentially arising confusion about the formalisms of Shannon Information and Kolmogorov Complexity. For the purpose of our following analysis of information in physics, we therefore have to provide straight forward presentations of the in this thesis examined information measures.

Step 2: Analyzing Information Measures

Only after completing *Step 1*, with a clear layout of the different information measures, we are able to proceed with *Step 2* of our ontological analysis. Such an analysis entails the following ingredients. On the one hand, we examine whether the supposed semantic/syntactic distinction indeed holds. On the other hand, the bearing to perceive Shannon Information (as many do) as a measure of uncertainty, has to be investigated. Then, we analyze to what extent information can be regarded as a concrete or abstract entity. Finally, we point out to what extent the different information measures are conventional. The reason for pointing out conventional elements, is that *entirely* conventional, so to speak 'made up' theories, can hardly qualify for an ontological (mind-) independent entity in the catalog of the world's furniture. After that, we're poised to work out the ontological status of the individual information measures and compare them among each other.

1.5 Thesis Contribution & Organisation

This section briefly describes the arrangement of this thesis' chapters and how to approach it. The main contribution of this thesis is to compare the notion of *algorithmic information* (*Kolmogorov Complexity*), so far underrepresented in foundation of physics literature, with the much more often analyzed *Shannon Information*. The purpose of such a comparison is to answer the main research question, i.e. to investigate the ontological status of information in physics using a yet largely neglected information measure. Note that answering this question doesn't imply to examine one of the many interesting information theoretical based approaches which try to explain (certain aspects about) the constitution of our world.

For so doing, we have to apply the just now developed framework. According to *Step 1*, we first need to have a clear notion of the different information measures. The following chapter, Chapter 2: *Measures of Information in Communication*, introduces the most important aspects

of Shannon's information measure. In the same vein, Chapter 3: *Kolmogorov Complexity–The Algorithmic Approach*, presents the formal features of algorithmic information. In the following chapter, Chapter 4: *A (Surprising) Connection between H(X) and K(x)* the formal connection between the previously introduced information measures (Shannon Information and Kolmogorov Complexity) is presented.

Thereafter, *Step 2* of our framework requires the 'ontological analysis' of our information measures and their connection, hence the following two chapters, Chapter 5: *Interpretation of Information in the Classical Case* and Chapter 6: *Interpreting Kolmogorov Complexity in the Classical Case*. Note that the latter also entails a comparison between Shannon Information and Kolmogorov Complexity in the classical case.

After having covered the classical case, we repeat the above structure for the quantum case. However, many of our insights from the classical case can be used as an auxilliary basis for the quantum case. Instead of the five chapters in the classical case, we only need two for the quantum case. Based on *Step 1*, Chapter 7: *Quantum Information Theory* introduces both Quantum Shannon Information and Quantum Kolmogorov Complexity. Based on *Step 2*, Chapter 8: *Quantum Information: What's the Ontology?* analyses both these quantum information measures in respect to their ontological status.

Finally, in Chapter 9: *Results & Outlook*, we conclude our thesis, based on the findings of the previous chapters. Finally, the reader might find useful additional information (in the semantic sense) in Chapter 10: *Appendix*.

Chapter 2

Measures of Information in Communication

"The word 'information' has been given different meanings by various writers in the general field of information theory. [...] It is hardly to be expected that a single concept of information would satisfactorily account for the numerous possible applications of this general field." [Shannon, 1993, p.180]

–Shannon

In the following chapter, we are presenting two classical information measures that arose in the context of classical communication theory–Hartley's combinatorial approach and Shannon's probabilistic approach. In foresight of the following chapters, we especially devote our attention to the probabilistic approach.

2.1 A Communication Model

For a better understanding of Shannon's information measure, it is instructive to familiarize oneself with a standard communication model. Shannon [Shannon, 1948] suggested a communication system to consist of five parts, as seen in Fig. (2.1). Of course, such a system was, as its very name suggests, meant to implemented in context of communication.

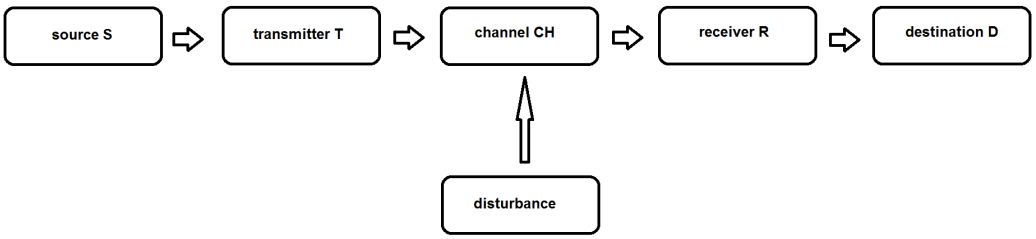

Figure 2.1: A communication model according to [Shannon, 1948]

1. S denotes the source. The information source produces a sequence of a set letters $\{x_i\}$, i.e. producing a message or a sequence of messages, which are to be communicated to the destination D.

2. T denotes the transmitter. The transmitter produces a signal, generated at S, to transmit in a suitable way for a channel CH (for instance, differences in sound pressure could be changed into an electric current).

3. CH denotes the channel. The channel is the medium to transmit the signal from T to a receiver R.

4. R denotes the receiver. The receiver reconstructs the message from the signal, hence performing the inverse operation of T.

5. D denotes the destination. The destination is the person or object for whom the message is intended, reproducing a set of letters $\{y_i\}$.

6. Furthermore, external disturbances have to be considered in case of a non-noiseless channel.

Note however, that there are basically no restrictions on what counts as a 'source' or any other part of the system. Such a model of a communication system doesn't even require the involvement of (conscious) agents. So we can, for instance, think of a communication system of two tin cans connected by a wire (transmission of the signal over sound through

the wire); two cellphones (transmission of the signal through electromagnetic waves); biological processes, where e.g. the DNA of reproducing cells is contained in daughter cells (transmission through molecules; (...).

It is this wide range of generality which allows for the successful application of $H(X)$ in virtually every scientific area. For the remainder of this thesis it is important to note that the above examples are exclusively cases of noisy communication channels. In the following though, we are not interested with the technical challenges to overcome noisy communication and only deal with the ideal noiseless case, if not stated otherwise.[1]

2.2 Hartley's Information Measure–The Combinatorial Approach

Even though often neglected, it is instructive to first briefly introduce the predecessor of Shannon's information measure–Hartley's combinatorial approach. Let in the following X denote a set of characters of the source S

$$X = \{x_i,\ i = 1, ...,\ n\}\,, \tag{2.1}$$

such that X can be, for instance, conceived as an alphabet like

$$X_{abc} = \{a,\ b,\ c, ...,\ x,\ y,\ z\}\,. \tag{2.2}$$

For a sequence of length N, one can select each of the n letters, N times, obtaining

$$W = n^N \tag{2.3}$$

as the number of possible sequences. As Hartley pointed out though [Hartley, 1928], for a measure of information to be of *practical* value in a context of engineering, it should be rather proportional to the number of selections N, instead of the number of possible sequences W (2.3). Since we usually measure magnitudes linearly, i.e. we ideally want information to be an extensive quantity like mass or volume, we obtain Hartley's

[1] N.B. a noiseless channel doesn't automatically guarantee successful communication. We shall deal with the so called 'success criteria' in section (5.5).

information measure

$$\log_2 W = NH_H, \tag{2.4}$$

$$\text{with } H_H = \log_2 n, \tag{2.5}$$

by applying the logarithm.

An example. At a given source S, Alice decides to send Bob at destination D two messages produced of the set of letters X_{abc} (2.2) with length $N = 5$, stating

$$m_1 = h\ e\ l\ l\ o, \text{ and}$$

$$m_2 = k\ j\ x\ g\ z.$$

According to eq. (2.5) in the ideal case of not taking into account noise, both of Alice's messages contain

$$NH_H(m_1) = NH_H(m_2) = 5 \cdot \log_2 26 \approx 23.5 \text{ bit} \tag{2.6}$$

of information.[2] By adding $H_H(m_1)$ and $H_H(m_2)$ we get approximately 47 bit of information and satisfy our intuitive notion to measure magnitudes linearly. Note additionally that the content of the messages is not relevant for the measure H_H and deprived of 'psychological considerations' [Hartley, 1928]. Both messages yield the same *quantity* of information. Following our claims in Chapter 1, we are only interested in syntactic information anyway, not semantic information.

However, even though we don't want to consider any potential meaning conveyed in the messages, we shouldn't neglect the different occurrences of letters. In the context of human communication, it doesn't come as a surprise for a person with the ability to understand English that certain messages and letters appear more often than others.

[2]Note that the base of the logarithm is in principle arbitrary, but standardly chosen to be two, because of the communication theoretic background of the Shannon theory. The unit of H is measured in binary digits, i.e. *bit* (in case of base 10 logarithms, the unit is called *hartley*, or *nat* when based on the natural logarithm). In the following, $'\log'$denotes the logarithm to base two, if not stated otherwise.

2.3 Shannon's Information Measure–The probabilistic approach

Let's regard our exemplary messages $m_1$and m_2 from the section above again. Human agents communicate in a certain language with specific letter frequencies, *effectively* leading to a probability distribution of the letters. In fact, the letters$\{k,\ j,\ x,\ g,\ z\}$ found in m_2 above are the least used ones in English and one would expect to find much more messages like m_1 with much more frequently used letters.[3] In contrast to Hartley, Shannon noticed that by exploiting the statistical properties of an *information source*[4], one can find a function $H(X)$, which always satisfies

$$NH(X) \leq N \log n. \tag{2.7}$$

where the right hand side of the inequality denotes NH_H, thus obtaining a more efficient way to transmit messages than with H_H. The basic idea behind this discovery is that due to non-identically distributed (i.e. based on a probability distribution $P(X)$) random variables, some sequences will be so unlikely (called *atypical*), that they can be left aside from our considerations from the start. Shannon's information measure hence determines how much a sequence can be *compressed on average* relative to simply enumerating all possible sequences. In fact, Shannon could demonstrate that $H(X)$ turns out to be the *optimal* statistical compression for using minimal physical resources for transmitting information. In general this result is known as *Shannon's noiseless coding theorem* [Shannon, 1948].

[3]Of course the frequencies and hence the effective probability distribution of the letters change in different languages. For a detailed overview of letter frequencies in English and German, see table 10.1 in the Appendix. For a more detailed and 'philosophical' analysis on how $H(X)$ depends on the underlying probability distribution P see Chapter 5.

[4]A classical information source can be modeled as a source that consists of a sequence of random variables x_i, where such variables are assumed to be *independent* and *identically distributed* (i.i.d.) [Nielsen and Chuang, 2000].

2.3.1 Intuitive derivation of Shannon Information

In [Shannon, 1948], Shannon originally chose the approach to derive $H(X)$ from a set of axioms. Since his derivation is far from straight forward and bears some interesting puzzles, we shall discuss his approach to some extent at a later point. For a deeper understanding of $H(X)$, it is instructive to only start with an intuitive derivation [Brukner and Zeilinger, 2001], [Timpson, 2013]. According to the *Law of Large Numbers* (*LLN*), the *average* of the outcomes obtained from a large number of trials N, ought to be close to their *expected value* (for sufficiently large N). One way of conceiving an i.i.d. information source S is then in analogue to an urn model (in which the balls are replaced). The different letters of the alphabet X are represented by the balls in the urn, which are distributed according to a probability distribution $P(X)$.

Based on the *LLN*, we then might consider an urn model *without* replacement to depict the idea of considering only so called $\varepsilon - typical$ sequences in the build up of our derivation of $H(X)$.[5] In that case, the length of the sequence has to be equivalent with the numbers of balls N in the urn. For demonstrating this, let's assume that we have a 'large' sequence of length–for the sake of the argument $N = 10^4$–with the English alphabet $X_{abc}$with its according letter frequencies (found in table (10.1), see Appendix), effectively giving us a probability distribution $P(X_{abc})$. For the letter 'z' for instance, we obtain an probability of $p_z = 0.0007$, so that we get seven '$z-$ balls' out of a total of 10^4 balls (made up of the balls for the rest of the letters). In such a model it is impossible to obtain an atypical sequence of thousand successive z's, as there aren't enough '$z-$ balls' in the urn to draw from. Ultimately, it is the *LLN*–by changing to an urn model without replacement–which implicitly allows us to restrict the number of '$z-$ balls' to only seven, as the average outcome should be arbitrarily close to the expected value (i.e. with large N, the probability distribution peaks around its mean value).

[5] N.B. one should take this 'intuitive derivation' with a grain of salt though, since an urn without replacement violates the assumption that the outcomes are independent from one another.

For a message of length N, where $N \to \infty$,

$$W_{typ} = \frac{N!}{Np(x_1)!...Np(x_n)!} \tag{2.8}$$

then enumerates the number of equiprobable typical sequences. Applying the Stirling approximation, we can simplify eq. (2.8) and obtain

$$W_{typ} = 2^{NH(X)}, \tag{2.9}$$

where

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log_2 p(x_i) \tag{2.10}$$

denotes the famous *Shannon Information*. The Shannon information thus quantifies the number of $\varepsilon -$ typical messages by $2^{NH(X)}$ (2.9) from a prearranged set of all possibilities $2^{N \log n}$. Each typical message can be *encoded* in a binary sequence by $NH(X)$ bits. Note that the encoding of each typical message into a binary sequence facilitates the transmission with digital technology; in practice, we often use differences in voltages to represent binary sequences. Generally speaking, the idea to achieve efficient coding is to assign shorter codewords to more frequent letters. Such a procedure may yield an optimal average encoding rate–$H(X)$. Once an encoded binary sequence reaches the receiver, it may be decoded for the destination. We'll go into a deeper analysis about coding in section (4.3).

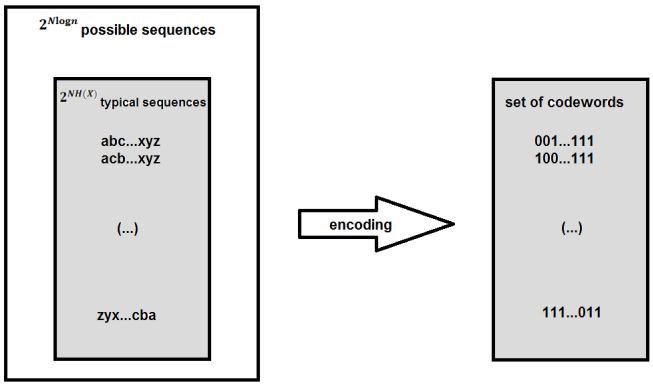

Figure 2.2: The subset of typical sequences $2^{NH(X)} \leq 2^{N \log n}$ is replaced with a binary code number of $NH(X)$ bits. Usually the encoded message is sent to the receiver.

2.3.2 Some mathematical properties of *H(X)*

Let's now discuss the most relevant *mathematical* properties of Shannon Information. For that purpose it's instructive to compare some of the features of $H(X)$ (2.10) with its historical predecessor, the Hartley Information H_H.

In the case of a binary source, i.e. a source with two outcomes (e.g. '1'and '0'), distributed according to probabilities p and $(1-p)$, $H(X)$ becomes

$$H_{\text{bin}}(p) = -p\log(p) - (1-p)\log(1-p). \tag{2.11}$$

$H_{\text{bin}}(p)$ is plotted in figure (2.3); the horizontal axes the probability p of one of the two outcomes and the vertical axes denotes the value of H_{bin}in bits, where the maximum value describes the case of no compression at all.

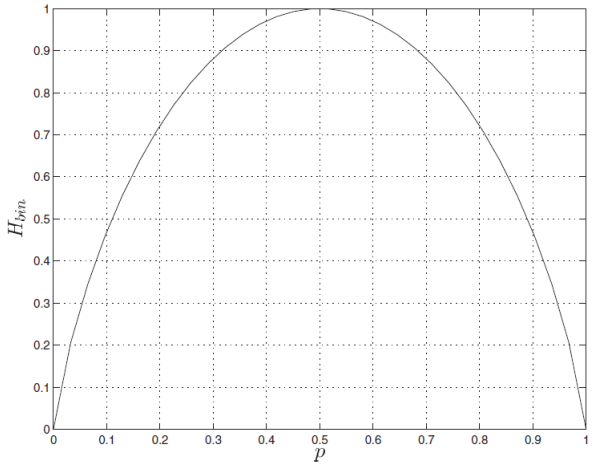

Figure 2.3: Binary entropy function for $H_{\text{bin}}(p) = -p\log(p) - (1-p)\log(1-p)$.

Analyzing the above figure, we can easily 'read off' that inequality $H(X) \leq H_H(X)$ (2.7) holds for the binary case. The function $H_{\text{bin}}(p)$ obtains its maximum value $H_{\text{bin max}}(p) = 1$ (in which case a sequence can't be compressed) under the condition that p and $(1-p)$ are equivalent, which for the binary alphabet is $p = \frac{1}{2}$. It is easy to show that in the case where all the probabilities of the outcomes of the source are equivalent, Shan-

non's Information measure becomes equivalent to Hartley's measure

$$H_{\max}(X) = H_H. \tag{2.12}$$

In other words, whenever the probabilities of the two outcomes are not equivalent, they are smaller than H_H, or otherwise equal.

In brevity, we can keep in mind, that in general $H(X)$ has the following properties

1. $H(X) \geq 0$

2. $H = 0$, if any only if all p_i equal zero, save but one having unity as value.

3. $H(X)$ reaches its maximum (and coincides with the Hartley measure $H_H = \log n$) when, for a given n, all the probabilities are equal $p_i = \frac{1}{n}$.[6]

4. $H(X)$ is a continuous function.

5. $H(X)$ *only* depends on the probability distribution $P(X)$ of the set of characters X of the information source S.

2.3.3 Joint-, Conditional- and Mutual Information

So far, we have only focused on the source as part of the communication system. However, in communication we are not only concerned about the set of characters X of the information source S, but also about the communication channel CH and its capacity C_{CH}, and the *decoded* set of letters Y at the Destination D.

[6]Note that in the context of statistical mechanics we find a somewhat similar discussion around the maximum entropy principle [Uffink, 1995]. For the reader with a background in physics, it might then be helpful to think of the somewhat analogous case of Boltzmann- and Gibbs entropy. In the usual notation, the Boltzmann entropy is given by $S_B = k_B \ln W$, where k_B denotes the Boltzmann constant and W the number of *possibilities* of phase points in a phase space Ω (such that $W \subseteq \Omega$).The Gibbs entropy on the other hand is given by $S_G = -k_B \sum p_i \ln p_i$, where p_i are the probabilities that a system (i.e. a collection of particles) lies in a certain region of Ω.

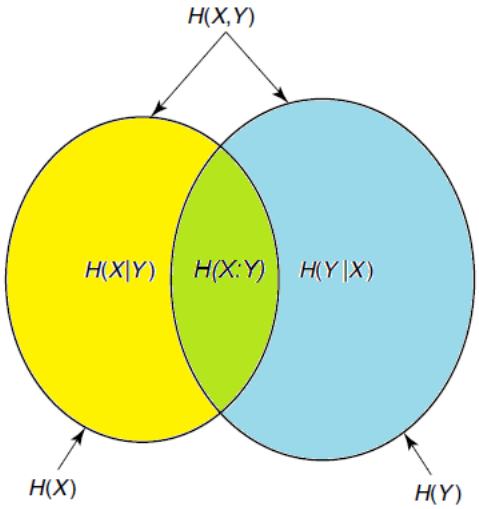

Figure 2.4: Diagram displaying Joint- $H(X,Y)$, Conditional- $H(X|Y)$, and Mutual Information $H(X:Y)$ [Cover and Thomas, 2006].

In the following, we will introduce the notions of Joint, Conditional, Mutual and relative Information. Their respective links are shown in Fig. (2.4); a diagram for the various information measures in respect to the correlated random variables X (the yellow and green area $H(X)$) and Y (the blue and green area $H(Y)$).

Joint Information *H(X,Y)*

Until now, we have only dealt with one random variable X. However, the notion of $H(X)$ can easily be extended to a pair of random variables, governed by a joint probability distribution $p(x,y)$

$$H(X,Y) = -\sum_x \sum_y p(x,y) \log p(x,y). \tag{2.13}$$

Expression (2.13) is called *joint information*. As illustrated in Fig. (2.4), Joint information $H(X,Y)$ can be interpreted as the combination of $H(X)$ and $H(Y)$, thus containing the yellow, green and blue area. Note however, that at first sight there is nothing surprisingly new in eq. (2.13), as (X,Y) could be simply considered a single vector-valued random variable. Yet,

the notion of Joint Information is a useful tool to get a grip on Conditional and Mutual information.

Conditional Information

The *conditional information*, also called *equivocation* of X given Y, describes the average amount of information generated at source S but *not* received at the destination D. With the so called *chain rule*, we can express such a relation as

$$H(X|Y) = H(X,Y) - H(Y), \tag{2.14}$$

linking conditional information to joint information (2.13). Regarding Fig. (2.4), $H(X|Y)$ is given by the yellow area.

In turn,

$$H(Y|X) = H(X,Y) - H(X) \tag{2.15}$$

is the *noise*, i.e. the average amount of information received at D but not generated at S. From Fig. (2.4) we can read off that the noise is depicted by the blue area.

Formally $H(X|Y)$ is then defined as

$$H(X|Y) = \sum p(y) \left(- \sum p(x|y) \log p(x|y) \right). \tag{2.16}$$

Another noteworthy property of the conditional information is that

$$H(X|Y) \leq H(X), \tag{2.17}$$

with equality if and only if X and Y are independent. We shall discuss the implications of eq. (2.17) in connection with the claim 'uncertainty can't increase' at full length in a later chapter. In terms of Fig. (2.4), statistical independence of X and Y would be given in the case, when $H(X)$ and $H(Y)$ had no overlap, hence not creating the green area.

Mutual Information

Mutual information on the other hand, is concerned with the question how much a random variable X can convey about a random variable Y. The mutual information is the *relative information*[7] between the joint distribution (2.13) and the product distribution $p(x)p(y)$ and given by

$$H(X:Y) = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}. \tag{2.18}$$

In Fig. (2.4), we find Mutual information to be denoted by the green area. In relation to the previous information measures, we find

$$H(X:Y) = H(X) - H(X|Y) \tag{2.19}$$

where $H(X|Y)$ is the conditional entropy of X given Y. With the chain rule (2.14) we obtain the following relation

$$H(X:Y) = H(X) + H(Y) - H(X,Y), \tag{2.20}$$

with the joint entropy $H(X,Y)$ of X and Y, as defined above. Furthermore, we can easily extract from Fig. (2.4), that

$$H(X:Y) = H(Y:X) \tag{2.21}$$

the symmetry of mutual information holds.

Finally, we are able to to define the *channel capacity* as

$$C_{CH} = \sup_{p(x_i)} H(X:Y), \tag{2.22}$$

where the supremum is taken over all possible input distributions $p(x_i)$.

The Fidelity function

With our extended notions of information at hand (2.13, 2.16, 2.18), we are able to shortly introduce the *fidelity* $\upsilon\left(P(X,Y)\right)$ as a measure of successful communication. The fidelity is a function of the joint probability of

[7] N.B. so far we were only concerned with *absolute* measures of information.

the source S and destination D and indicates the necessary resources to communicate a message. In case of a *continuous* probability distribution governing the source, Shannon [Shannon, 1948] suggested the fidelity to be

$$\upsilon\left(P(X,Y)\right) = \int\int p(x,y)\rho(x,y)\mathrm{d}x\mathrm{d}y, \tag{2.23}$$

where$\rho(x,y)$ denotes how desirable it is to receive y when x was sent. However, Shannon didn't specify the details of $\rho(x,y)$, leaving the interpretation of successful communication to us. We shall discuss these matters in Chapter 5.

2.4 Summary

We introduced the structure of communication systems and the most common formal aspects of information measures stemming from communication theory. Even though Hartley's combinatorial-approach is hardly paid attention to today, it served as an insightful introduction to its successor–the probabilistic-approach. The latter allowed us to introduce Shannon's noiseless coding theorem, which specifies the minimal resources for noiseless communication. Instead of regarding all combinatorial possible messages, we only make allowance for probable, so called $\varepsilon-$typical sequences, which are subsequently encoded into $NH(X)$ bits. The Shannon Information $H(X)$ is a measure of the minimal resources (the average number of bits per symbol) needed to reliably encode the output of an information source. Additionally, we introduced further formal repertoire of Shannon's theorem, for instance, specifying conditional information or the channel capacity, some of which shall be discussed in Chapter 5.

Chapter 3

Kolmogorov Complexity - The Algorithmic Approach

"But what real meaning is there, for example, in asking how much information is contained in "War and Peace"? Is it reasonable to include this novel in the set of "possible novels," or even to postulate some probability distribution for this set? Or, on the other hand, must we assume that the individual scenes in this book form a random sequence with "stochastic relations" that damp out quite rapidly over a distance of several pages?" [Kolmogorov, 1965, p.6]

– Kolmogorov

In the previous chapter, we encountered Hartley's combinatorial approach and Shannon's statistical approach to measuring information. However, in contrast to both of these information measures, which are only sensible in context of an *ensemble* of messages, we focus on the algorithmic approach to measure information of *single* sequences in this section.

For different reasons and motivations, Algorithmic Information Theory (AIT), was at first independently articulated by various scholars in the 1960s (see Chapter 1). Nowadays the algorithmic information measure is often referred to as *Kolmogorov Complexity*. In general, *computational complexities* (i.e. the analysis of algorithms), are either concerned with determining the amount of time, storage and/or other resources to

execute algorithms on computers. Kolmogorov Complexity on the other hand, denotes the length l of the shortest program p that generates a particular sequence x (on a Turing Machine)

$$K(x) := \min l(p), \tag{3.1}$$

measured in bits.

3.1 Intuitive derivation of Kolmogorov Complexity

Let's first look at a rather intuitive 'derivation' of Kolmogorov Complexity. Introducing the algorithmic approach in his paper [Kolmogorov, 1965], Kolmogorov wondered (see epigraph), about the information content of individual sequences. His questions have to be conceived in contrast to Shannon's theory which (as we have seen in the previous sections) is explicitly *not* concerned with the meaning of single messages and only dependents on a given probability distribution. In a sense, Kolmogorov's questions are related to the *infinite monkey theorem*. The theorem is based on the idea, that after waiting long enough, a monkey *randomly* hitting the keys of a typewriter, will with certainty reproduce any given text.[1] Intuitively however, it seems indeed puzzling that sequences with clear *patterns*, such as novels like *War and Peace* for instance, should be regarded as merely one of an arbitrary ensemble of sequences. The optimality of Shannon's noiseless coding theorem with respect to the average message may be not optimal for individual cases at all.

For solving such a puzzling 'uneasiness', we need a method to determine the *randomness* of sequences. In the case of a random statistical process such as coin tossing e.g., all the potentially produced sequences happen to be equally likely. Regarding the following sequences[2]

[1] Of course, the 'monkey randomly hitting the typewriter' is just an amusing way to represent an i.i.d. information source. If real life monkeys were able to be trained to typewrite, the messages would probably show patterns.

[2] Remember, the fact that the sequences are binary ones, doesn't have to concern us. As will be explained in a section on Coding Theory (4.1), we can in principle encode any

$$x_1 = 10101010101010101010101010101010,$$

$$x_2 = 11000001011010010111001100111110,$$

both sequences have indeed the same probability $p(x_1) = p(x_2) = \left(\frac{1}{2}\right)^{30}$ to occur. Ordinary information theory though, offers no solution for calling strings of length N 'more random' than another. Hence, our inability to compare individual sequences in respect of their randomness, is based on the inherently statistical and probability based conception of Shannon information. In order to overcome such an inability, we have to examine single sequences.

Shifting our focus on *individual* strings, in fact allows us to introduce a notion of randomness. Glancing at the sequences $x_1$and x_2, it should be immediately apparent, that the former one is purely repetitive and can be simply described as $N\ times\ '10'$, with in our case $N = 15$. Following this method of describing strings, we just discovered, that some strings can be compressed considerably. In case of such a compression, we say that x_1 is a *simple* or *regular* sequence. However, for structureless sequences of the kind like x_2, i.e. strings without any patterns, such a form of compression is not possible.[3] Instead, we need a lot more effort to specify x_2 and in the worst case, the shortest description is just the sequence itself. In case of no (noteworthy) compression, we call a string *(Kolmogorov) random*.[4]

However, we have to be careful about self-contradictory descriptions like the *Berry Paradox*, where we define

> "[t]he least natural number that cannot be described in fewer than twenty words." [Li and Vitanyi, 2008, p.1]

In case the described number exists, we have just described it in thirteen words, contradicting its own definition. Therefore, we should be well advised to only accept descriptions that are explicit enough for giving us

given text (including *War and Peace*) into a binary code.

[3]N.b., it is possible, that string the $x_2$might indeed have some patterns we are not immediately aware of though. A more thoroughly discussion about the procedure of deciding whether a string is in fact random, will follow in chapter 6.

[4]In fact, $x_2$was generated by 'randomly'tossing a coin 30 times

instructions in order to construct the corresponding string in an unambiguous and purely mechanical manner [Mueller, 2007]. For doing so, it has been proven very helpful to make use of computers (or rather the formal notion of Turing Machines), conceiving the description of sequences as algorithms which let a predefined computer halt and (within a finite amount of time) put out some string. The usage of computers doesn't only force us to use the just stipulated precise descriptions or algorithms, but also serves as a formal procedure to determine the randomness of a string.

3.2 Turing Machines

For formalizing algorithmic information, it is instructive to first discuss the underlying notion of *Turing Machines* ($\mathcal{TM}$). Additionally, we have to agree on the definition of a reference model for our computations of algorithmic complexity. For that purpose, we shall especially have a closer look at the subclass of the so called *Universal Turing Machines* ($\mathcal{UTM}$).

In 1937, Alan Turing investigated what it means for a task to be computable. He wondered whether the thoughts of a human (brain) could be equally processed by an inanimate device. Formalizing the intuitive idea about how humans usually try to solve a task, they tend to think, write, think again, etc., Turing came up with an effective procedure, called *algorithm*, to describe a set of instructions for such a device, nowadays called Turing Machine.

The set up of a $\mathcal{TM}$ can be imagined to be constructed of three basic mechanical components (see Fig. 3.1).

1. The first component is a two-way infinite *tape* divided into cells. Each of those cells contains a symbol from a finite set of *input symbols* Σ or the *blank* symbol b. For reasons of convenience, we in the following assume that Σ only contains the symbols $\{0, 1\}$. In that case, the input tape contains blank symbols b in each cell, save for the finite number of cells that hold $\{0, 1\}$, i.e. the input symbols Σ. Both Σ and b are contained in Γ, the complete set of *tape symbols*;

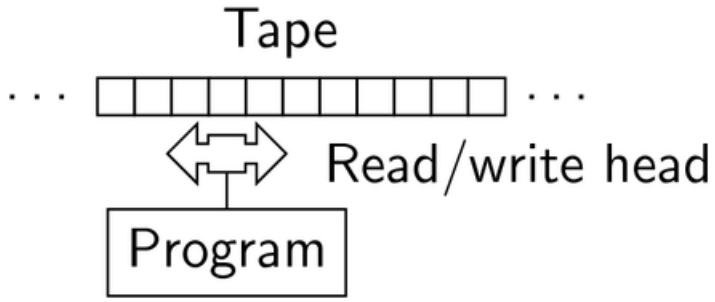

Figure 3.1: Schematic Turing machine

Σ is always a subset of Γ.

2. The machine requires a *read-write head* for scanning the cells on the tape. The read-write head is able to move to the left L and right R along the tape to scan its successive cells, one at a time. At each step, the $\mathcal{TM}$ is in one of the *control states* Q, including the *initial state* q_0 and the subset $\mathcal{F}$ of several *halting states*.

3. At last, a *table of commands* contains a set of transition rules, which determine the action of the machine in the next cycle (in a way the table can be regarded as 'program' of the $\mathcal{TM}$). A cycle is defined as the transition from one state of the machine into another, where the $\mathcal{TM}$ can then take the following actions: moving its head to the left or right, and erase or write in the scanned cell. If a given input doesn't yield a final or *halting state*, the $\mathcal{TM}$ may continue to operate forever in the above described cycles. Each transition can be described by the so called transition function δ, defined by the next control state Q, a symbol (in Γ) being replaced from the tape, and the direction in which the head moves (L or R).

While the above description focused on an actual physical implementation, we have to keep in mind that $\mathcal{TM}$s don't require to be 'real'physical objects, in fact they are mathematical models or *abstract* machines. Formally, the 7-tuple of the previously defined components

$$M = \{Q, \Sigma, \Gamma, \delta, q_0, b, \mathcal{F}\} \tag{3.2}$$

then defines a $\mathcal{TM}$ [John E. Hopcroft and Ullman, 2001].

Let us now briefly discuss what a $\mathcal{TM}$ can and cannot do (for convenience, we will in the following only regard $\mathcal{TM}$'s that read the input tape from left to right). According to the *Church-Turing Thesis*, every effective computation can be carried out by a $\mathcal{TM}$ [Copeland, 2015].With the above described principles, we are then able to compute all arithmetic and logical functions; all functions computable on a $\mathcal{TM}$ are *recursive functions*.

However, known as the *Halting problem,* it is generally not possible to predict whether or not a given input will lead to a halting computation. In other words, a function $h(t, n)$ that would determine whether a machine t halts on input n, is a *non*-recursive function (N.B. the computability of recursive functions and the halting problem will be examined in Chapter 6 in context of the analysis of algorithmic information). Applied to real programs and stated in less technical terms, there is no computer program that examines the code for a program and determines whether that program halts [Barker-Plummer, 2016].

Universal Turing Machines ($\mathcal{UTM}$)

The above description allows us to extent the idea to *Universal Turing Machines* ($\mathcal{UTM}$), which are capable of simulating the behavior of any other $\mathcal{TM}$. Since $\mathcal{TM}$s are constructed from three basic components (i.e., a tape, a read-write head, and a table of commands), the only way in which the output any two $\mathcal{TM}$s can differ, is in the initial configuration of the tape, the internal state of the finite state control, and the table of commands.The idea behind $\mathcal{UTM}$s is to fix the table of commands and the finite state control, such that the initial contents of the input tape are left as the only variable. The universal machine then reads both the description of the $\mathcal{TM}$ to be simulated as well as the input from such a $\mathcal{TM}$from its own tape.

3.3 Formal presentation of *K(x)*

The Kolmogorov Complexity of a finite sequence x is defined as a function from *binary* strings of arbitrary length to the natural numbers

$$K : \{0,1\} \rightarrow \mathbb{N}. \tag{3.3}$$

With the short introduction in the previous section about $\mathcal{TM}$s, we can now have a closer look at the formal details of Algorithmic Information Theory. Instead of some arbitrary description method p, we now conceive our description as an algorithm or program operating on a $\mathcal{TM}$,

$$p(y) = x, \tag{3.4}$$

which with input y, prints x and halts. Considering the preliminary notion (3.1) we've introduced earlier, we now define Kolmogorov Complexity

$$K_{\mathcal{U}}(x) \ := \min \, l(p) + l(y), \tag{3.5}$$

where $l(p)$ denotes the length of the program p, and $l(y)$ the length of the program y (both represented in bits) [Gruenwald and Vitanyi, 2008].

With the definition (3.5) above, we seemingly defined a *non-unique* information measure though, as $K(x)$ is both dependent on a certain program language and a reference machine $\mathcal{U}$. In the following paragraphs (*Invariance Theorem* and *Universality*), we show that $K(x)$ in fact represents an *absolute* and '*objective*' information measure.

Invariance Theorem

We want to draw attention to the choice of a suited description method p (3.4), which has to be encoded in a certain arbitrarily chosen program language.[5] However, the *Invariance Theorem* shows (see e.g.,[Devine, 2009]) that save for a constant $O(1)$, $K(x)$ does not depend on the used program language , such that

[5] We're going to deepen our understanding about Coding Theory in the following Chapter. The notion of D is made precise in (4.4).

$$K_1(x) \leq K_2(x) + O(1), \tag{3.6}$$

where $K_1(x)$ and $K_2(x)$ are Kolmogorov complexities, each formulated in different program languages.[6] In other words, changing from one program language into another, only changes the value of Kolmogorov Complexity (of sufficiently long sequences) up to a fixed constant $O(1)$.

Universality

Next to the coding scheme (i.e. in fact the program language), $K(x)$ also depends on the $\mathcal{UTM}$ on which it is (arbitrarily) chosen to be calculated on. In similar vein to the Invariance Theorem above though, it can be demonstrated that (given the reference machine $\mathcal{U}$),

$$K_{\mathcal{U}}(x) \leq K_{\mathcal{A}}(x) + c_{\mathcal{A}}, \tag{3.7}$$

where $K_{\mathcal{U}}(x)$ and $K_{\mathcal{A}}(x)$ denote the complexities calculated by different $\mathcal{UTM}$s [Cover and Thomas, 2006]. The constant $c_{\mathcal{A}}$ on the other hand, corresponds to the size of the program that will simulate one $\mathcal{UTM}$ on another one. Even though, there exists an infinite number of $\mathcal{UTM}$s, we can choose a reference $\mathcal{UTM}$ '$\mathcal{U}$' with a certain instruction set, such that $c_{\mathcal{A}}$ can be made (negligibly) small relative to the actual description of a long string. Up to a constant $c_{\mathcal{A}}$, our choice of a specific reference machine is irrelevant, such that we drop the subscript '$\mathcal{U}$' from here on.

Complexities of Regular and Random Sequences

Let us now take a closer look at the complexity of regular and random strings once again. Earlier, we explained the concept, that simple or regular strings display a certain pattern and can be compressed

$$K(x) = O(\log_2 n). \tag{3.8}$$

[6] Standard programming languages for calculating $K(x)$ on actual computers are for instance, LISP or Java.

However, as can be proven, the overwhelming majority of n-bit sequences are in fact random. Since there can't be many strings having *unique* a short description, the majority of strings will have to rely on a long description (i.e. algorithm p). It is easy to see that for every *uniquely decodable* code, there are no more than 2^n strings for which x can be described in n bits, because there a simply no more than 2^n binary strings of length n [Gruenwald and Vitanyi, 2008]. This amounts to $2^{n-1} + 2^{n-2} + ... + 1 < 2^n$ as the number of strings which can be *at most* described by less than n bits. The fractions of strings with $K(x) < n - k$ is smaller than 2^{-k}. Accordingly, programs whose description are for instance only 10 bits shorter than the actual length n of the sequence (so a $n - 10$ bit algorithm), can account for at most only for 2^{-10} or $1/1024$ of all the n-bit strings.

For completely random sequences we then obtain a scaling of $n + O(1)$, where $O(1)$ denotes a constant that accounts for the commands such as *print* and *halt*. Yet, for technical reasons,[7] we are from now only interested in the so-called prefix complexity, i.e. an information measure for which the input to the associated $\mathcal{TM}$ results in a halting computation that is prefix free.[8]

3.3.1 Joint, Conditional and Mutual Complexity

Analogue to the case of Shannon's information measure, we are able to construct a couple of information measures derived from the algorithmic information content $K(x)$. Since the ideas behind joint complexity, conditional complexity and mutual complexities are in fact very similar to section (2.3.3), the explanation will be somewhat shorter here. For a better understanding the reader may go to the respective section regarding $H(X)$.

Nevertheless, we have to make some adjustments, to account for the

[7] The main reason is that the here defined prefix complexity, simplifies the comparison with Shannon information in the following. Note that there exists a version of Kolmogorov Complexity which doesn't necessarily depend on prefix-free programs, but such a version won't be discussed here.

[8] We'll discuss the notion of prefix free codes in the next chapter in detail.

considerations made around Universality and the Invariance Theorem. In analogue to [Gruenwald and Vitanyi, 2008], we introduce the following notation; for an inequality within an additive constant, we write '$<^+$', so that when f and g are functions from $\{0,1\}$to $\mathbb{R}$, '$f(x) <^+ g(x)$' means $f(x) < g(x) + c$. In addition, '$=^+$' means $f(x) = g(x) + c$.

Joint Complexity

The *joint complexity* $K(x, y)$, is defined as the size of the smallest program of two sequences x and y, calculating them simultaneously, one has

$$K(x, y) \leq K(x) + K(y) + O(1). \tag{3.9}$$

In the case where equality holds (in fact '$=^+$'), we call two strings *algorithmically independent*, so that there is no algorithm p that is capable of computing both x *and* y, shorter than stringing the programs which individually compute x and y together. [Desurvire, 2009].

Conditional Complexity

Relative or *conditional complexity* $K(x|y)$ of x given y, is defined to be the size of the smallest program to calculate x from a minimal program for y

$$K(x|y) =^+ K(x, y) - K(y). \tag{3.10}$$

In the case where x and y are algorithmically independent (there is no help from x to compute y) $K(x|y) = K(x)$ and likewise $K(y|x) = K(y)$. With the notions of joint complexity and conditional complexity, we can now introduce mutual complexity.

Mutual Information

Mutual complexity $K(x : y)$ measures the commonality of x and y, i.e. it ascribes a value to the extent to which knowing y helps one to calculate x

$$K(x : y) = K(x) + K(y) - K(x, y) =^+ K(x) - K(x|y). \tag{3.11}$$

For two strings that are *algorithmically independent*, their mutual complexity is virtually zero (or more precisely, as small as possible). Considering two arbitrary strings x and y of length N, knowing one of them usually won't help calculating the other. In most cases x and y will be random to each other.

3.3.2 Uncomputability of *K(x)*

Let us now take a closer look at how to determine Kolmogorov Complexity. As we have seen in (3.2), the Halting Problem states that certain functions are not computable. Unfortunately, $K(x)$ it is not a computable function; given a string x it is impossible to find an algorithm which exactly determines the shortest program p to compute x. Put differently, $K(x)$ does not fall into the class of recursive functions.

The full implications of the uncomputability will be discussed at a later point (see section (6.1)). As the last formal aspect, we want to point out that the uncomputability of $K(x)$ doesn't make it a completely useless notion. For practical purposes $K(x)$ can often be approximated.

Methods of Approximation

A simple procedure to approximate $K(x)$ may follow these steps (given a string x and a TM):

1. Let $\mathcal{TM}$ generate a lexicographic list of all input programs p. Such a list may look like $p : \{0,1\} = \{b, 0, 1, 00, 01, 10, 11, 000, 001, ...\}$, where b denotes the blank symbol;

2. Let $\mathcal{TM}$ run all lexicographic ordered programs in parallel. In the first cycle, $\mathcal{TM}$ runs the blank program b; in the second cycle b and 0; in the third $b, 0$, the successive program and so on. By computing all programs in parallel, it is guaranteed that the $\mathcal{TM}$ won't stop with the procedure when a halting program is computed.

3. Let $\mathcal{TM}$ record a list of all halting programs;

4. Let $\mathcal{TM}$ check for all of these halting programs, whether the resulting output string is x;

Eventually, the estimate of the Kolmogorov Complexity is given by the shortest program that has resulted in putting out x.

3.4 Summary

Algorithmic information, often also referred to as Kolmogorov Complexity $K(x)$, is based on the idea that the descriptions of objects may display regularities and can be compressed. Such descriptions are always conceived as sequences or strings x, and the more patterns there are, the shorter these strings will be. In order to formalize the idea, these descriptive sequences can be calculated by algorithms p operating on a universal Turing machine $\mathcal{UTM}$. $K(x)$ is then defined as the length of the shortest algorithm p, yielding x. In addition, the introduction of Kolmogorov Complexity provides us with an intuitive notion of randomness. In fact, most sequences are patternless and therefore random. However, because of the Halting problem, we can't always compute $K(x)$ and thus know if a sequence is in fact random or not.

Chapter 4

A (surprising) connection between *H(X)* and *K(x)*

> "“[...] when I did discover algorithmic probability, I realized that it was the inverse of Huffman's problem. He obtained a short code from knowledge of probabilities. I obtained probabilities from knowledge of short codes.” [Solomonoff, 1997, p. 76]
>
> – *Solomonoff*

In this chapter, we point out the formal connection between Shannon's Information measure and algorithmic information, i.e. Kolmogorov Complexity. At first, such a connection perhaps seems surprising, as both approaches are be based on different notions. Whereas Shannon's Information measure was originally developed in the context of communication theory and denotes the statistical compressibility of an ensemble of possible messages (which is in addition often held for a degree of uncertainty), Kolmogorov Complexity measures the minimal length of a program operating on an $\mathcal{UTM}$, generating a single string.

However, as we'll see in the following, both approaches have a close connection through Coding Theory. In addition, we examine the so called *Universal Probability* and establish a connection between $K(x)$ and $H(X)$. Thereafter, we will show yet another way (next to the intuitive and axiomatic approaches) to derive Shannon's information measure through

Coding Theory. In the end of this section, we will then demonstrate that

$$\sum_x p(x)K(x) \approx H(X) \tag{4.1}$$

holds.

4.1 Coding Theory

For the following sections it is instructive to have a short introduction to Coding Theory, as both Shannon's theory and Kolmogorov Complexity rely on *prefix free coding*.[1]

One of the fundamental concerns of Coding Theory is the search for efficient ways to represent a set of symbols $X = \{x_1, x_2, ..., x_n\}$, put out by a kind of source with the respective probability distribution $P = \{p_1, p_2, ..., p_n\}$, in terms of another set of codewords $C = \{c_1, c_2, ..., c_n\}$. For all practical purposes, we require efficiency in coding, i.e. the idea of compressing information by assigning short descriptions to the most frequent outcomes of X and longer descriptions to less frequent ones. For the reason of convenient storage and transmission, the code words of C are often composed from a binary alphabet, according to standard notated as $\{0, 1\}$, such that

$$X \to C. \tag{4.2}$$

A well known example is Morse Coding, where the set of symbols X is the standard English alphabet and the letter 'e', for instance–the most frequent letter in the English alphabet (see Appendic (10.1))–is encoded by a single dot $\bullet$ (one of the codewords of C).[2] By following the method of assigning longer codewords c_i to less occurring letters ('i' e.g., is encoded by $\bullet\bullet$) we obtain *variable-length* codewords,[3] with the *mean codeword length*

[1]The actual term is 'prefix coding', but 'prefix free coding' is more intuitive and increasingly used nowadays.

[2]In the case of Morse Code$\{0, 1\}$are notated as $\{\bullet, -\}$.

[3]For sake of brevity, we only deal with variable-length codewords in this section. Of course one can also use fixed-length codes for coding.

$$L(X) = \sum_i p_i l_i \tag{4.3}$$

with probability p_i and codeword length l_i.

However, while we achieved that every *individual* codeword is unique, a sequence of such codewords is not necessarily unique either. Typically though, we want an unambiguous procedure D_f (the *decoding function*) to decode an encoded message, such that

$$D_f(c_i) = x_i. \tag{4.4}$$

Considering for instance, the short Morse Code

$$\bullet\,\bullet\bullet, \tag{4.5}$$

we are faced with various non-unique possibilities for decoding that sequence. Without further specifications, we are unable to decide between $\{e, e, e\}, \{e, i\}, \{i, e\}$and $\{s\}$ as the 'right' decoded sequence.

The issues we are confronted with in our example, are not only constrained to Morse Code, but to a wider class of so called non-*prefix (free) codes*. The problem of such codes is that we can't specify where one codeword ends and another one begins, i.e. some codewords are a prefix of another codeword. Regarding Morse Code, the problem is circumvented by introducing '/' as the an extra sign [Desurvire, 2009], so that by adding '/' to equation (4.5) we obtain e.g.,

$$\bullet/\,\bullet\,\bullet,$$

now being *uniquely decodable* as $\{e, i\}$.

However, by choosing the strategy of adding an extra symbol (*merely* acting a space holder), we lose a lot of *coding efficiency*. In virtually every context though, we want a highly efficient code–often solely based on two symbols (as pointed out above). For avoiding the defects of inefficient and non-prefix free codes, we have to optimize our coding efficiency.

There are several methods for optimizing codes (e.g., *Shannon-Fano*

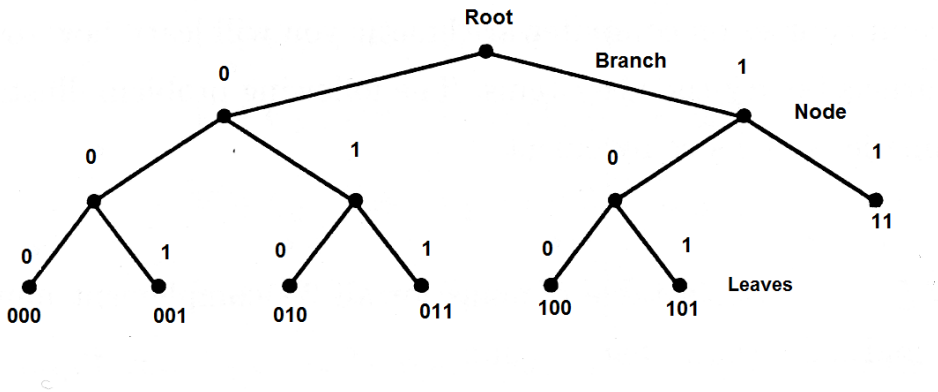

Figure 4.1: A binary code tree with seven code words.

Coding or *Huffmann Coding*). Roughly, the idea for finite codes is that the optimality is governed by how closely the length of the set of code words C relates to the probability distribution of the set of source words [Li and Vitanyi, 2008]. For our purpose, the idea for efficient codes and prefix free codes can be visualized in a *code tree* as seen in Fig. (4.1). The pictured code tree is a binary one; starting from the so called 'root' (the beginning of the tree), one can choose to follow one of the 'branches' of the code tree, repeating the decision at every 'node', until reaching a leaf. For uniquely generated codewords, one can consider codewords generated at the nodes and the leaves (as in the case of Morse Code). However, by choosing *only* leaves as possible options for the codewords, we obtain unique *and* prefix free codewords.

For a binary tree as depicted above, with n leaves, for n codewords $C = \{c_1, c_2, ..., c_n\}$ with various lengths l_i, the *Kraft Inequality*

$$\sum_{i \in C}^{n} 2^{-l_i} \leq 1 \tag{4.6}$$

must hold.[4] In the depicted code tree above (Fig. 4.1), we have seven codewords; six codewords with $l = 3$ (where the length can also be determined by counting the number of branches one has to follow to arrive at

[4]For proving the Kraft Inequality, one has to sort the codewords by their length, such that $l_n \geq l_{n-1} \geq ... \geq l_1$. A binary tree may only have 2^{l_n} leaves, which number reduces by $2^{l_n - li}$ for every codeword of length l_i, such that $2^{l_n} - \sum_{i=1}^{n} \frac{2^{l_n}}{2^{l_i}} \geq 0$. Rewriting the last expression we obtain eq. (4.6).

a leaf) and one codeword with $l = 2$, such that $6 \cdot \frac{1}{8} + \frac{1}{4} = 1$ and the Kraft inequality holds.

For an efficient code (or optimal code), we have to reduce the mean codeword length $L(X)$ (4.3) as much as possible. We call a code *complete*, if the addition of any new codeword to the set of codewords C yields a non-uniquely decodable code. Substituting the summand of the Kraft Inequality (4.6) with

$$p_i = 2^{-l_i}, \tag{4.7}$$

results in $\sum_i p_i \leq 1$, with equality for complete codes. In regard to Kolmogorov's axioms of probability [Kolmogorov, 1933] (see also section (5.1)), all the requirements of the axioms are met, so that we can now interpret the p_i as *probabilities* with respect to some probability distribution P [Gruenwald and Vitanyi, 2004]. By remodeling (4.7) and taking into account that l_i is an integer, we then obtain the following lower bound restriction[5]

$$l_i \geq -\log_2 p_i, \tag{4.8}$$

for the length of the codeword, based on the probability distribution P of the codewords. Put differently, more frequent codewords should be shorter than less frequent ones, i.e. 'the leaf should be closer to the root'.

4.2 Universal Probability

In [Solomonoff, 1964], Solomonoff developed a quantitative formal theory of induction based on $\mathcal{UTM}$s and algorithmic complexity. For solving problems regarding probabilistic models of induction he developed the notion of *algorithmic probability* (also more commonly referred to as *Universal Probability*).[6] So far, we introduced algorithmic complexity based

[5]Because the codeword lengths must be integers, the optimal condition $l_i = -\log p_i$ can merely be approximately satisfied in the general case. The codeword length then has be chosen as the greater than but closest to l_i. Such code assignment is also known as *Shannon-Fano code* [Desurvire, 2009].

[6]The basic idea of Solomonoff's idea is that nature follows some unknown computable probability distribution.

on the idea that simple or regular sequences can be described by a much shorter algorithm than their actual length n. The majority of sequences is random though, where finding such a short algorithm is impossible. However, instead of starting with a given sequence x and asking how much it can be compressed by finding an algorithm p describing it, we can also begin 'from the other end', wondering which computable process produces a string x in the first place. Formally, such a computable process is based on a program p, which itself is a string of $\{0, 1\}$ when regarded as the input of a $\mathcal{UTM}$.

For illustration of Solomonoff's concept, let's assume that we're generating some arbitrary sequences p of $\{0, 1\}$ by e.g., noting the outcomes from flipping a fair coin. Providing an $\mathcal{UTM}$ with our arbitrarily generated sequence p as input, may result in the machine to halt and put out a particular sequence x. We now take all input sequences, i.e. programs p, that uniquely produce the output sequence x of the $\mathcal{UTM}$, and define the algorithmic probability of x as

$$m(x) = \sum_{p:\ \mathcal{U}(p)=x} 2^{-l(p)} \leq 1, \tag{4.9}$$

where $l(p)$ is the length of the input sequence (the program p). The right hand side of equation (4.9) thus denotes the probability to obtain a sequence x (as output of a $\mathcal{UTM}$) by a randomly generated program (of length $l(p)$). Note that $m(x)$ simply equals the left hand side the Kraft Inequality (4.6) we've introduced in the section about Coding Theory (4.1) in regard of uniquely decodable codes. However, as we are overall only interested in prefix free complexity, there is *only one* sequence or program p which generates our target sequence x, such that we don't have to sum over several programs. In addition, we demand p to be the shortest of all such prefix free sequences. With (3.1), we can simply identify $\min l(p)$ as the definition of $K(x)$, such that

$$m(x) = 2^{-K(x)} \tag{4.10}$$

denotes the algorithmic- or universal probability [Vitanyi, 2012]. It's im-

portant to note, that $m(x)$ merely represents a probability, not a probability distribution. Even by summing over the algorithmic probabilities, we only achieve $\sum_x m(x) \leq 1$, denoting as so called semimeasure (instead of a probability distribution with $\sum_x p(x) = 1$).

By solving for $K(x)$, we then obtain

$$-\log_2 m(x) = K(x), \tag{4.11}$$

known as the *Coding Theorem*. The left hand side of the above result (4.11) strikingly reminds us of the logarithmic part $\log p(x_i)$ of Shannon's information measure $H(X)$ (2.10). It is high time, that we turn to investigate the relation between Shannon Information and said Coding Theory.

4.3 Linking Shannon Information with Coding

So far we haven't explicitly pointed out how $H(X)$ accounts for the encoding of messages. Let's recall Shannon's Noiseless Coding Theorem; in a nutshell it establishes the lower bound for compressing messages to such a rate, at which in a perfect communication scenario (i.e. the case for a noiseless channel) no information will be lost. $H(X)$ thus describes the most efficient way of coding the set of typical messages, as only communicating with an optimal code guarantees lossless communication with minimal channel capacity $C_{CH} = \sup_{p(x_i)} H(X : Y)$ (2.22).[7]

From the section about Coding Theory, we remember that the goal for an optimal code is to find the lowest mean average code word length $L(X) = \sum p_i l_i$ (4.3). We can thus conjecture that every non-optimal code has

$$L(X) \geq H(X). \tag{4.12}$$

Since we assume the various probabilities p_i of the individual codewords

[7] In our case of an ideal noiseless channel $H(X : Y) = H(X)$ holds, such that $\sup H(X)$ determines the at least required channel capacity C_{CH}. Since $H(X)$ is the optimal encoding function, every other encoding is suboptimal and requires a greater channel capacity.

c_i to be simply given, we only have to be concerned with finding the respective minimal codeword lenghts l_i. By reformulating the Kraft Inequality (see above), we obtained the lower bound restriction for the codeword lengths $l_i \geq -\log_2 p_i$ (4.8); plugging this restriction into $L(X)$ and assuming equality $l_i = -\log p_i$, we get

$$\min L(X) = -\sum p_i \log_2 p_i. \tag{4.13}$$

With the inequality between $L(X)$ and $H(X)$ (4.12), we can conclude that equality holds for

$$\min L(X) = H(X),$$

$$\min \sum p_i l_i = -\sum p_i \log_2 p_i, \tag{4.14}$$

giving us *Shannon's information measure as minimal average codeword length.*[8]

4.4 Formal connection

As we have seen, Kolmogorov Complexity, which we originally introduced as the shortest *encoded* description of an object, can also be introduced on the basis of algorithmic *probability*. Vice versa, the *probability* based notion of Shannon Information can also be understood as the minimal average *code*word length. In addition, we already pointed out, that the Kolmogorov Complexity $K(x) = -\log m(x)$ (see (4.11)), based on the universal distribution $m(x)$, strikingly looks like the logarithmic term '$-\log p_i$' of $H(X)$ (2.10). Since $K(x)$ is the minimal *individual* description length, we can wonder whether by multiplying $K(x)$ (4.11) with the probability distribution $P(X)$ found in $H(X)$, we can achieve $\sum p(x)K(x) \approx H(X)$, such that it equals the minimal *average* code word length.

[8]For a more rigorous derivation of $H(X)$ as the minimal average codeword length with the method of Lagrange multipliers, see Appendix (10.2).

From (4.12) we can deduce

$$\sum p(x)K(x) \geq H(X). \tag{4.15}$$

Furthermore, it can then be shown in a more formal manner that the *expected complexity* $\sum p(x)K(x)$ is indeed asymptotically equal to $H(X)$, such that

$$0 \leq \left(\sum_x P(x)K(x) - H(P) \right) \leq c_p, \tag{4.16}$$

with $c_p = K(P) + O(1)$ (a constant which only depends on P) and $H(P) = -\sum_x P(x) \log P(x)$, with the important condition that P is a *computable* probability function. Since the exact proof would exceed the limits of this thesis, we refer to [Cover and Thomas, 2006, §14.3] or [Li and Vitanyi, 2008, Theorem 8.1.1].

Conditional, Joint and Mutual Information

In the chapters on Shannon Information and Kolmogorov Complexity, we encountered the conditional, joint, and mutual information, respectively complexity measures. Regarding these measures, Li and Vitanyi conclude that

> "[f]or almost every Shannon theory notion there turns out to be a Kolmogorov complexity theory notion that is equivalent in the sense that the expectation of the latter is closely related to the former."[Li and Vitanyi, 2008, p. 603]

In fact, not only do $H(X)$ and $K(x)$ virtually obtain the same value for very long sequences, but moreover, also each of their conditional, joint and mutual versions do. In the following we decided to omit explicitly showing the exact relations between conditional, joint and mutual information/complexity since the additional benefits for the goal of this thesis are marginal. The reader who might be interested beyond that, shall be advised to look at [Gruenwald and Vitanyi, 2004], [Li and Vitanyi, 2008] and [Harremoes and Topsoe, 2008].

4.5 An example - Algorithmic entropy

Since the connection between $K(x)$ and $H(X)$ seems largely underrepresented in physics so far, there aren't many examples where applying the above connection between weighed average of Kolmogorov Complexity and Shannon Information goes beyond mere theoretical considerations like eq. (4.16). One of the only areas of application in physics can be found in relation to computation (approached by the insights of statistical mechanics and Thermodynamics). In such a context, we can indeed find a few examples in [Zurek, 1989], [Li and Vitanyi, 1992] (and to some extent [Bennett, 1982]) discussing the limits of *Maxwell's Demon*.[9] Before starting to illustrate one of said examples, a few important remarks we have to keep in mind. Since the following example is based on statistical mechanics and Thermodynamics, we will unavoidably be involved in the discussions around entropy. In a sense, this is a rather unfortunate choice (there aren't better alternatives though), as 'entropy' is yet another heavily overworked term in context of information theoretic talk (remember the quote in Chapter 1, where Shannon and von Neumann discussed how to call $H(X)$). The exact relation between entropy and information can't be dealt with in a short section like this and deserves to be treated in an independent thesis. One should therefore be very careful as taking this section as the suggestion that 'information is entropy' or vice versa.

Derivation of the Sackur-Tetrode equation with K(x)

Let's now consider a randomly distributed mono-atomic gas of N particles, contained in an isolated box with volume V at temperature T. Our task is to find is to find a short algorithm that describes a classically mono-atomic gas. The location of a single particle of such a gas shall be determined with an accuracy of $\Delta V = \Delta_x^D$, where D denotes the dimension of the box. Here ΔV is the cell volume and the process of localizing a particle can be regarded as the analogue to *coarse graining* as classi-

[9] This section is partly based on a section of tutorial paper in 'Foundations of Statistical Mechanics and Thermodynamics' (Utrecht University, academic year 2016/2017), the author of this thesis wrote about the comparison of statistical and algorithmic entropy.

cally done in statistical mechanics. Encoding the location of each particle scales with $\sim \log_2 \frac{V}{\Delta V}$, where $\frac{V}{\Delta V}$ denotes the *number* of cells (remember, encoding an integer N with a binary code scales with $\log_2 N$). Moreover, the location of each of the N particles has to be described by D integers, giving us an additional $\sim \log_2(ND)$ term, such that

$$K \simeq N \log_2(\frac{V}{\Delta V}) + O(\log_2(ND)) + O(1). \tag{4.17}$$

Because the term $O(\log_2(ND))$ poses only a small correction, it can be incorporated as part of the constant $O(1)$. In addition, considering the indistinguishably of the N particles of the mono-atomic gas, we can even further compress K, by supplying only the differences of the particles' location instead of their individual locations. Based on that Zurek demonstrates [Zurek, 1989] that we get an extra $\frac{1}{N}$ term in the logarithm $K \simeq N \log_2(\frac{V}{N\Delta V}) + O(1)$.

Moreover, a program is needed to encode the momentum of the individual particles. As in the case just described for the coordinates, we obtain a term $\log_2 \frac{p}{\Delta_p}$, where the expected value of p is $(mk_{\mathrm{B}}T)^{\frac{1}{2}}$. The size of a typical program to encode all the momentum components for one particle is then given by $D \log_2 \frac{(mk_{\mathrm{B}}T)^{\frac{1}{2}}}{\Delta_p}$. Finally, putting our considerations about the coordinates and momenta of the particles together, one obtains

$$K = N \left(\log_2 \frac{V}{N\Delta V} + \frac{D}{2} \log_2 \frac{mk_{\mathrm{B}}T}{(\Delta_p)^2} \right) + O(1), \tag{4.18}$$

which is formally identical to the *Sackur-Tetrode equation*[10] expressing the entropy of the gas.

4.6 Summary

At first sight conceptually different information measures, Shannon's theory and Kolmogorov Complexity, both offer notions of probability and coding based interpretations. The noiseless coding theorem allows us to construe that Shannon Information is the optimal lower bound for coding

[10] For a detailed summary of the derivation of the Sackur-Tetrode equation in classical context, see [Grimus, 2011].

(i.e., the minimal average code word length). This code is entirely based on the probabilistic characteristics P_S of the source and doesn't use any features of the encoded object or message itself. Solomonoff's thoughts about inductive inference then introduced us to algorithmic probability and the Universal Probability $m(x)$. By then solely describing the characteristics of an (individual) object, we obtain a string x from which we can construct a code that's independent of P_S but yields an optimal codeword length too; we identified Kolmogorov Complexity $K(x)$ as exactly this kind of code. Any differences between those two encoding schemes for a particular sequence x is then based on the different starting points of those codes–exploiting the probabilistic behavior of the source versus the patterns in x. The two code word lengths coincide, when we consider a probability distribution that accounts for the regularities in x [Li and Vitanyi, 2008]. As we've seen earlier, the universal probability $m(x)$ of x meets these requirements.

Chapter 5

Interpreting Shannon Information in the Classical Case

> "The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point. Frequently these messages have meaning [...]. These semantic aspects of communication are irrelevant to the engineering problem."
> [Shannon, 1948, p. 410]
> – *Shannon*

BEFORE looking at the information measures in the quantum case and being able to examine the ontological status of information therein, it is highly instructive to interpret the formal mathematical frameworks of Shannon information (Chapter 2), Kolmogorov Complexity (Chapter 3) and their formal relation (Chapter 4) in the classical case. As we shall see later, many of these interpretations will help us to figure out the ontological status of information in the quantum case. This chapter begins by analyzing Shannon Information in the classical case.

5.1 Untangling Uncertainty

We start by examining the relationship of Shannon Information with uncertainty. Understanding this relationship is crucial for our judgement of the ontological status of Shannon information.

Next to the intuitive derivation (Chapter 2) and Shannon's axiomatic derivation,[1] we can derive $H(X)$ as a *measure of uncertainty*. Remember, Shannon himself was in doubt whether his information measure should be referred to as uncertainty. Intuitively, the link between uncertainty and information might be conceived as 'less uncertainty equals more information.' Often it's stated that a more concentrated probability distribution $P_S(X)$ of the alphabet of the information source yields a better *prediction* or 'less surprise' of the next letter. In addition, $NH(X)$ can be regarded as the number of yes or no questions one has to ask for revealing which binary sequence was sent. A less concentrated probability distribution yields a higher value of $H(X)$ which translates into having to ask more questions. Having to ask more questions can then be interpreted as being more uncertain. However, before going too deep into these kind of discussions, let's first ask if Shannon information really equates with uncertainty.

For our following analysis, we first require a more sophisticated notion of uncertainty which makes it essential to be aware of the distinction between the concept of uncertainty in prediction and inference. While the former is tightly related to the concept of *probability*, the latter is connected to the idea of *likelihood*. In our everyday language, 'probability' and 'likelihood' are often almost used interchangeable.[2] Whether we may say "It is likely going to rain tomorrow" or "It is probably going to rain tomorrow" won't make much of a difference, unless one is utterly nitpicking. In a formal sense though, the concepts of likelihood and probability are fundamentally different concepts–$H(X)$ can only be conceived as a measure of uncertainty in prediction.

[1] For the sake of completeness, the interested reader is invited to consult the Appendix (10.3) in order to compare (Shannon's) axiomatic derivation of $H(X)$ with the following uncertainty based derivation.

[2] At least this seems to be the case in English.

Likelihood - Uncertainty in Inference

For reasons of illustration we are confronted with an urn experiment. We don't know anything about the content of the urn, i.e. we are clueless about the number of balls (or if there are any) and their labeling. Despite the mysteriousness of the urn, we are brave enough to draw objects from the urn and observe a certain number of outcomes. For clarity, we assume that six red balls and four blue balls have been drawn (replacing the balls and shaking the urn after each draw). Now we wonder with what (un)certainty we can *infer* the underlying probability distribution (i.e. the ratio between the balls in the urn) to be 0.6 (red balls) to 0.4 (blue balls). Stepping aside from the urn example, the problem generally runs down to our uncertainty about what can be *inferred* from a given sample. An information measure concerned with inference is, for instance *Fisher Information* [Fisher, 1925].

Probability - Uncertainty in Prediction

Let's regard our urn-example for a second time.[3] Curious as we are, we peeked inside the urn and saw that it was filled with red and blue balls only. Before drawing a ball, we can then *predict* the outcome with a certain *probability* based on the ratio of the two colors. In the case of an equal amount of red and blue balls, we are maximally uncertain about which color we obtain from drawing one ball. In reverse, when *it is known* that for instance, only a few red and a lot of blue balls are placed in the urn, then we are fairly certain to draw a blue ball. In other words, intuitively our uncertainty is high when the probability distribution about the outcome is spread; our uncertainty is low when such a probability distribution is concentrated. Instead of uncertainty we can then equally well speak about the *concentration* of probability distributions.

Comparing the probability density functions $p(x)$ and $q(x)$ shown in Fig. (5.1) for instance, we can intuitively tell that the former is more concentrated than the latter, hence displaying less uncertainty. [4]

[3] In fact, a very similar example can originally be found in [Hilgevoord and Uffink, 1991] and [Uffink, 1991]

[4] However, to account for a formal definition of uncertainty is not a trivial task and

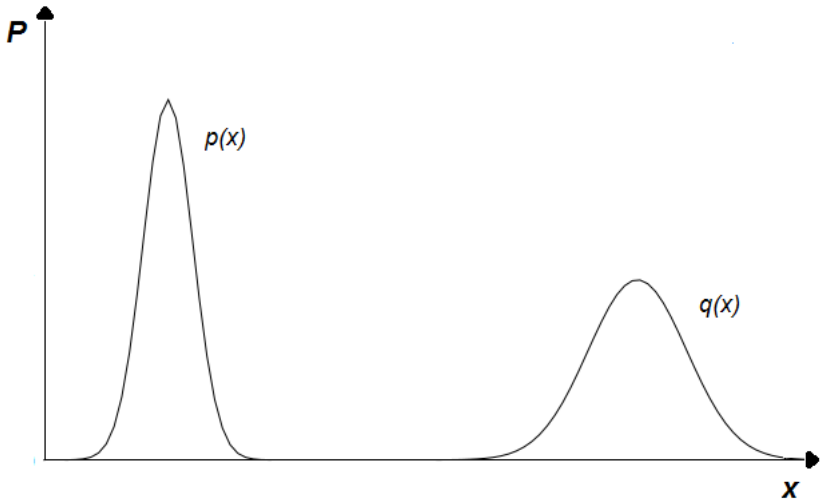

Figure 5.1: Two probability distributions $p(x)$ and $q(x)$.

Uncertainty in communication scenarios

Now equipped with a notion of uncertainty in prediction, we can apply the concept to the probability distribution of an information source. Often it's then *very* useful to think of the Shannon information $H(X)$ not only as 'statistical compressibility' but also as a measure of uncertainty.

From now on we may regard the above filled urn as an information source S emitting a sequence of red and blue balls. For an evenly distributed probability distribution, we are maximally uncertain to predict what the next letter (in our case 'ball') might be. In case of an highly uneven distributed probability distribution, our uncertainty about the next outcome reduces. In the extreme case where the probability for a specific outcome equals one, there is no uncertainty at all such that $H(X) = 0$.

With a decreased uncertainty about individual letters, we also obtain a notion for the uncertainty of whole messages; being less uncertain about the outcome of a specific message, means having to ask less yes or no questions which sequence was sent.Instead of having to ask N yes

one has to overcome a series of challenges (e.g., how to determine the concentration of a probability distribution with various 'spikes' or how to deal with conventional aspects like scaling the graphs).

or no questions, we now merely have to ask $NH(X)$ questions to reveal which sequence was sent.[5] As explained earlier (compare Chapter 2), applying the *LLN* to unevenly distributed probability distribution, reduces the average length of encoded messages, which requires less yes or no questions such that the uncertainty is decreased.

Even though regarding Shannon Information as uncertainty seems to be intuitively appealing and very useful in many practical applications, we shall point out parts of Timpson's *deflationary view* (cf.[Timpson, 2004], [Timpson, 2013]) that in the following prohibits us to equate uncertainty and Shannon information when it comes to ontology.

Uncertainty based derivation of $H(X)$

In his dissertation [Uffink, 1991], Jos Uffink defines a general class of measures of uncertainty $U_r(P, \mu)$, where P is a probability distribution and μ a background measure. Demanding that any uncertainty measure $U_r(P, \mu)$ ought to be

1. invariant under permutations;

2. continuous; and

3. strictly Schur concave;

he concludes that

$$U(P, \mu) = \chi^{-1}\left(\sum \mu_i \phi(\frac{p_i}{\mu_i})\right), \tag{5.1}$$

is the only expression which satisfies the above postulates (with χ being a continuous decreasing function, ϕ a convex function, and p_i probability measures).

With two additional scaling conventions, Uffink demonstrates that

$$H_r(P, \mu) = \log U_r(P, \mu), \tag{5.2}$$

[5]The strategy to reveal which binary sequence was sent, is to ask: 'Was the first letter a 1?' and itertate the question for all of the N letters of the sequence.

with $U_r(P, \mu)$ being the result of the additional scaling conventions [Uffink, 1991]. For $r = 0$ we get

$$H_0 = -\sum p_i \log \frac{p_i}{\mu_i}, \tag{5.3}$$

obtaining equivalence with Shannon's information measure (2.10) when μ denotes the counting measure, such that $\forall i : \ \mu_i = \#_i = 1$.

Having arrived at expression (5.3), thus means that $H(X)$ can be derived as an uncertainty based measure. Notice though, that such a derivation is *not unique*; equivalence with Shannon's measure *only* holds for $r = 0$, so that in any other case, i.e. $r \neq 0$, we obtain some other measure of uncertainty. We can thus conclude that $H(X)$ is merely one of many possible measures of uncertainty and falls into Uffink's general expressions of type U_r.

In order to derive Shannon's $H(X)$ uniquely, yet another axiom is required (let's call it *updating requirement*). Shannon requires that

> "The uncertainty about y is never increased by knowledge of x. It will be decreased unless x and y are independent events, in which case it is not changed.",

such that

$$U_r(P) \leq U_r(P*), \tag{5.4}$$

the updated uncertainty measure $U_r(P)$ (P denotes the 'updated' probability distribution due to our knowledge of x) is less than or equal to the original one $U_r(P*)$. Since it can be shown that expression (5.4) only holds for $r = 0$, we could then uniquely derive Shannon Information as a measure of uncertainty. By applying the additional scaling conventions (5.2) and (5.3) the updating requirement can be rewritten as

$$H(Y|X) \leq H(Y), \tag{5.5}$$

(compare (2.17)).

Where Shannon's requirement seems plausible in the beginning, the uncertainty about one variable can in fact be increased by knowledge of

the other. An example is provided by [Aczeel and Daroczy, 1975]: Ornitologists state the existence of white ravens, but of course the chances of spotting the next raven being white are really small; let's assume the probability to find a white one is 0.01 vs 0.99 to find a black one. In other words, our uncertainty about the color of the next raven is really small. In addition, the ornitologists found out, that if a raven has a white mother, the probability of the offspring being white is, let's say 0.5. By learning that a white raven mother has offspring, our uncertainty about the color of the offspring (i.e. the conditional information measure for one event) has *increased*; the probability distribution is less concentrated and has changed from 0.01 vs 0.99 to 0.5 vs 0.5.[6]

Counter examples as the one above, show that Shannon's justification for the updating requirement isn't sound, such that the requirement is untenable as a further axiom to uniquely pick out $H(X)$ as a measure of uncertainty. It's worthwhile to note that the raven example doesn't violate expression (5.5) though, for it merely holds *on average*. $H(Y|X)$ doesn't display the information of one particular distribution, but instead denotes the *expected value* of conditional information

$$H(Y|X) = \sum p(x) \left(- \sum p(x|y) \log p(x|y) \right).$$

As such, $H(Y|X)$ doesn't prohibit that the information of a conditional distribution over x is increased by an observation y in an individual case.

After all, we then did *not* proceeded to derive $H(X)$ as *as a unique* measure. As we shall see in the next subsection, the latter part of this diagnosis is very important for our following analysis.

[6] Uffink's suggests the 'keys in the pocket' example, picked up by [Timpson, 2004], where information about x may increase the uncertainty of x itself. With a high probability (so with little uncertainty) Uffink's keys are in his pocket, but if he discovers that they're not, then the probability distribution of the location of the keys becomes spread and the uncertainty of the whereabouts of the keys increase dramatically.

5.2 One Formalism–Many Interpretations

We have to clarify which formalism of $H(X)$ shall act as the basis for our ontological analysis. Based on the previous chapters and sections we have introduced the most important aspects for quantifying information, and by so doing, we have encountered at least two different interpretations of Shannon's information measure:

1. In context of communication, $H(X)$ is a measure of the *statistical compressibility* of a source, specifying the number of $\varepsilon -$ typical sequences $2^{NH(X)+\varepsilon}$ from all possible sequences. Instead of all possible messages ($N \log n$ bits), we merely have to encode each of the typical sequences with $NH(X)$ bits. According to Shannon's noiseless Coding Theorem, such a compression is optimal, such that $H(X)$ has to be a lower bound for the the minimum average code word length $\min L(X)$ eq. (4.3).

2. On the other hand, as pointed out in the former section about uncertainty (5.1), $H(X)$ displays a measure of uncertainty in prediction, i.e. the lack of concentration of a probability distribution. On this view, greater uncertainty yields a larger value of $H(X)$ and thus more yes or no questions which have to be asked in order to reveal which sequence was indeed sent.

Originally, $H(X)$ was introduced to solve the problem of communication, i.e. to discover which minimal channel capacity is needed to transmit all $\varepsilon -$ typical $2^{NH(X)+\varepsilon}$ messages from a pool of all possible $2^{N \log n}$ messages, emitted by a specific source S. The Shannon Coding theorem then states that $H(X)$ *uniquely* specifies the optimal rate of transmission of information. It is crucial to note that $H(X)$ in the context of communication, is a characterization of that specific source S, not any specific message. Such characterization of the source in turn, is *only* dependent on the respective probability distribution $P_S(X)$, which denotes the probabilities over the set of possible output states X of the source.

On the other hand, we may conceive $H(X)$ as a measure of uncertainty. And indeed, as demonstrated earlier, $H(X)$ is *one* special case

among many measures of uncertainties, all generalized by Uffink's $U_r(P)$. For many purposes, it is then *very useful* to think of $H(X)$ determining of the number of yes or no questions we have to ask for revealing a message.

So which Interpretation do we analyze?

For our following analysis it is decisive to point out which of the above interpretations we're going to look at. Whereas Shannon's Noiseless Coding Theorem states that $H(X)$ is a *unique* information measure, we concluded that the uncertainty based derivation of Shannon Information is unable to pick out $H(X)$ as a unique measure of uncertainty. The latter result thus stands at odds with Shannon's Coding Theorem; $H(X)$ can't be unique and non-unique at the same time! As a measure of Shannon information $H(X)$ is distinct from $H(X)$ as a measure of uncertainty. Thinking about $H(X)$ in terms of uncertainty is admittedly often very useful, but it appears to be pure happenstance that the equations for these different concepts coincide. In order to figure out the ontological status of information we must resist the view to interpret $H(X)$ as a measure of uncertainty. Instead we regard $H(X)$ as denoting the amount of Shannon Information an information source is able to produce.

5.3 No place for semantic elements

Let's continue by disburden our chosen interpretation of Shannon information from having anything to do with semantic information. After all, we still want to retain the syntactic/semantic distinction of information in physics we established in Chapter 1 (so far, we have merely claimed that such a distinction exists). For so doing, we return to the question whether labeled balls from an urn (or in fact sequences of such), count as information. Clearly, one might answer such a question with "look, there is a letter painted on this-and-that ball, that's information!". But we mustn't allow to ignore the semantic/syntactic distinction and relapse into thinking about information in physics in terms of meaning. For whatever historical reason and only because of mere *convention*, we use the symbols

'$\{a, b, c, ...\}$' for the letters '$\{a, b, c, ...\}$' in English [Duwell, 2008].

In order to escape the drawbacks of 'meaning' and to avoid the conventions of the English language, we can simply modify our urn example; by exchanging the balls labeled with Latin letters X_{abc} with some other set of balls labeled with an alphabet of 26 different 'letters', i.e. arbitrary but *distinguishable* symbols. In fact, we could as well fill our urn with, for instance, fruits and vegetables such that our new set of output states is characterized by the alphabet[7]

$$X_{F\&V} = \{Apple, Broccoli, Celery, ..., Zucchini\}\,. \tag{5.6}$$

Virtually any kinds of tokens can act as an alphabet and as long as the new alphabet is still distributed to the prior probability distribution, the value of $H(X)$ remains unaffected. Changing the possible outcomes of the information source–including the unusual case of fruit and vegetables–ought to make obvious that the remark about English letters and (semantic) information were premature. Shannon's formalism is silent on the nature of possible output states and doesn't depend on whether these states are itself semantically relevant. As we've argued, virtually any distinguishable *tokens* the information source is able to emit, suffice to create a sequence/message. Claiming that all these possible kinds of tokens are per se semantically relevant seems an untenable position. After all, the nature of tokens the sequence consists of (e.g., the new set of balls or any other arbitrary set of output states), doesn't affect whether a message is meaningful or not.

However, the above argument adumbrated that we can encode messages in multiple ways by using many different kinds of tokens, without necessarily changing the value of (syntactic) information $H(X)$. In general, messages are quite independent of the tokens they're transmitted with, as long as these tokens follow a certain *syntax* that allows to encode (and successfully decode) the messages. Since the relation of the tokens is determined by what's encoded, one could therefore wonder if $H(X)$ se-

[7] Of course the initials of the fruits and vegetables don't have to match with the English alphabet for the argument to hold.

mantically accounts for such an encoded message. And surely, we can encode something meaningful in, let's say, a binary sequence. But trying to establish a connection between semantic information and $H(X)$ is untenable, as Shannon Information only depends on the probability distribution of the source. How should a probability distribution give rise to meaning? Such a probability distribution merely allows us to exploit certain (syntactical) regularities for optimal communication (i.e., only considering likely $\varepsilon -$ typical), not to quantify anything meaningful.

5.4 A Two Way Strategy

The former section raised the issue that in principle all kinds of tokens can be used to solve Shannon's 'fundamental problem of communication'. Does it therefore make sense to equate all these kinds of tokens with syntactic information? We've to take a closer look at the qualitative and quantitative aspects of $H(X)$ to answer this question.

Anew, we are confronted with an urn (filled with $N = 10^4$ balls) which are distributed according to $P(X_{abc})$–a probability distribution based on the letter frequencies in English. In this case, the urn acts as the information source, producing sequences of letters according to the drawn balls. As explained in Chapter 2, we're then able to calculate $H(X)$ according to (2.10)

$$NH(P(X_{abc})) = -10^4 \cdot (p_a \log p_a + p_b \log p_b + ... + p_z \log p_z)\,,$$

for a message of length $N = 10^4$. While this calculation *quantifies* a value for the amount of syntactic information measured in bits, it leaves us quite clueless of *what* is actually produced at the source; does it in this concrete case then make sense to speak of balls with letters on them as pieces of information?

Let's first look at the quantitative aspect. In general, from being able to quantify phenomena or entities with a certain formalism, a clear specification of what's in fact measured, doesn't necessarily follow.[8] Histori-

[8]For a more general treatment of these kind of problems, one might consider reading

cally, we may for instance draw the comparison to electric current

$$I = \frac{\mathrm{d}Q}{\mathrm{d}t}, \tag{5.7}$$

which is the flow of electric charge Q through a given surface over time t. At the time of the 'creation' of the formalism the view of what was actually flowing around, was certainly different from today, where we usually think of discrete charged particles (electrons or ions mostly). Without further context, equation (5.7) doesn't specify what's actually flowing around.

Considerations in the vein of the example above can be applied to the Shannon formalism, which in turn make it essential to distinguish between *quantity-information* and *type-information*.[9] Whereas the former notion quantifies an amount of information (in the senses seen above), only the latter (and for us much more interesting case) specifies what's actually 'produced' at the information source.

Quantity-Information

So while replacing the tokens used in communication doesn't change the quantity of $H(X)$ if the underlying probability distribution remains in fact the same, the value of Shannon information might change without ever changing the tokens in the set up (i.e., in case the underlying probability distribution changes). This suggests that quantity information is indeed quite independent of whatever tokens the source emits. Quantity information shall then be described as:

> *Quantity Information is a function dependent on the probability distribution of the source $P_S(X)$, which neither depends nor quantifies which kind of tokens the source uses in a communication scenario.*

[Frigg and Nguyen, 2016] as an introduction into the topic of scientific representation.

[9] To the authors knowledge, Timpson [Timpson, 2004] (see also his later writings) in his *deflationary view* was the first to suggested a distinction between quantity and pieces of information. Duwell [Duwell, 2008] later adopted this position and used the terms 'quantity-information' and 'type-information'.

Just like in the case in the case of the electric current I, we don't think of the quantity per se as something 'physical' or 'concrete'; such quantities or values are always abstracta. So while we can certainly spatio-temporally locate things like apples, chairs, and tigers, we can't do so with quantities. Quantity information merely provides a quantitative measure *of what* is on average *produced* at the information source. Quantity Information is hence an abstractum too.

Type-Information

What's emitted by an information source on the other hand, are the tokens that the specific source is capable of producing tokens of. Considering Shannon's quote in the epigraph of this chapter, we remember that the 'fundamental problem of communication' is the reproduction of a message (i.e., the output of a source) at the destination. As we've seen in the former section, many different types of tokens suffice to communicate messages; English letters, fruit and vegetables, smoke signals, sound, etc.. These concrete tokens merely represent the means for sending the *multiply realizable* message that is encoded according to certain (syntactical) rules determined by the minimum average codeword length $H(X)$, such that they can be reproduced (i.e., decoded) at the destination.[10]

Based on similar insights Timpson, then defined Type Information as:

> "$Information_t$ *is what is produced by an* $information_t$ *source that is required to be reproducible at the destination if the transmission is to be counted a success.*" [Timpson, 2013, p. 22][11]

Note that Timpson's definition doesn't rely on any particular sequence of tokens. What has to be identified to be reproducible is Type Information

[10] In fact, basically any every-day communication scenario involving digital technology involves encoding of the original message into a set of different tokens. For instance, while Alice and Bob are having a conversation on the phone, the message–an utterance based on sound waves (the 'original' tokens)–is converted into a binary electric signal, which Bob's telephone might successfully reproduce into the original message.

[11] Note that Timpson's $information_t$ can be understood as our concept of syntactic information (although he appears to disagree with using 'syntactic' in this context (see fn. 25 [Timpson, 2013])).

(a sequence type) which for each communication scenario, is based on a sequence of certain kinds of tokens.

Emphasizing our Coding Interpretation we might slightly adjust Timpson's quote by exchanging 'produced' and 'reproducible' by 'encoded' and 'decodable' respectively, and adding the role of tokens, such that our definition reads:

> *Information is what is encoded by some kind of tokens of an information source that is required to be decodable at the destination if the transmission is to be counted a success.*

By then examining *what is produced at the source*–Type Information, a particular kind of sequence types–we come to the conclusion (as the coinage of the term already suggests) that this entity is an *abstract type*.[12] Even though the source 'spits out' *concrete tokens*, these tokens are used to generate a signal which *encodes* the Type Information created by the source. What's important is that the multi realizable sequence type can be (according to some success criterion) reproduced at the destination. Virtually any kind of distinguishable tokens suffice to encode and a decode a sequence. In other words, particular sequences of concrete tokens then *instantiate* abstract type information, just like an apple or a cherry might instantiate the abstract notion of redness.

5.5 To what Extent is Shannon Information Conventional?

Having illuminated the distinction between Quantity- and Type Information, one might has gotten suspicious about conventional elements here and there, when, for instance, replacing labeled balls with fruits and vegetables. Does this conventional freedom bear on the ontological status of information? In addition, it's not obvious how, in the above definition of Type Information, to determine what actually counts as a success.

[12] In case one might have a different about *the problem of universals* and deny the existence of universal types, one can probably tell a similar story with *abstract* tropes.

Let us therefore first have a look at the conventional aspects of standard probability theory. Thereafter, we have to examine the success criteria of Shannon's 'fundamental problem of communication' and our earlier given definition of information.

The experimental set up

In *Grundbegriffe der Wahrscheinlichkeitsrechnung* [Kolmogorov, 1933], Kolmogorov formulated an axiomatic system that enables us to deal with probability distributions on discrete and continuous sets. In the standard formulation, a probability space is a triple $(E, \mathcal{F}, P)$, where i)E is a set, ii) a subset $\mathcal{F}$ (called σ-field) of E, and iii) the probability measure P. The point of departure in Kolmogorov's system is a set of *elementare Ereignisse* (*elementary e*vents) in E, to which probabilities are applied. Both probabilities and of events are assumed to be primitive [Galavotti, 2005]. In the style of our urn-example in the discussion around uncertainties (5.1), the elementary events in E can be taken as 'drawing a red ball' and 'drawing a blue ball'.[13] Since we are only interested in well-defined events, we regard the subset $\mathcal{F}$ of E, which is closed under union, (denumerable) intersections and complements. Probability is then introduced as a measure P on $\mathcal{F}$ which assigns a numerical value to each subset A_i of $\mathcal{F}$, in the sense that

1. (Non-Negativity)$P(A) \geq 1$ for all A in $\mathcal{F}$
2. (Finite additivity)$P\left(\cup_i A_i\right) = \sum_i P(A_i)$ if all A_i are mutually disjoint
3. (Normalization)$P(E) = 1$.

In order to apply Kolmogorov's formalism to our present case, it is useful to introduce the distinction between an *experimental set-up* and an *experiment* Uffink used in [Uffink, 1991]. We assume that an experimental set up is denoted by the triple $< E, \mathcal{F}, P >$, such that any conceivable

[13] In his original paper, Kolmogorov actually used the example of tossing a coin twice. The set of elementary elements then correspond to $\{Heads - Heads,\ Heads - Tails\ , Tails - Heads\ , Tails - Tails\}$. Provided with a sufficient amount of red and blue balls, the example can of course be easily transferred to our urn-example.

outcome counts as an elementary event. However, in an experiment we only regard those events which we're able to or *choose* to distinguish on a trial. Hence, an experiment is an experimental set-up for which the kind of trial is specified. For the formal description of an experiment, we then have to specify the quadruple $< E, \mathcal{F}, P, X >$, where X accounts for the partition of E according our distinction of the outcomes.

With a certain interpretive flexibility we can conceive a communication scenario as an experiment; in particular we mean each time an information source 'creates' a token as output. Returning to our paradigm example–the urn–drawing a ball from such is treated as the output of an information source and can be perceived as an experiment. The conventional element in such an experiment is then imported by choosing what counts as a different outcome; i.e., how the partition X of the given experiment looks like. For instance, do we treat the the following instances

$$\{A, \mathbf{A}, \boldsymbol{A}, \mathsf{A}, \mathtt{A}, \mathbf{A}, \mathfrak{A}, \mathcal{A}, \mathscr{A}, \} \tag{5.8}$$

all as outcomes of a ball with the letter 'A' or do we distinguish them any further? Alternatively, for some reason, we could only be interested in the occurrence of vowels and consonants, or merely distinguish between fruits and vegetables, instead of each individual member of the alphabet $\{Apple, Broccoli, ...\}$. In any case, choosing a partition X is completely stipulated by the conventional choices of the user and influences the probability distribution of the source $P_S(X)$.

So while we can partition at will, the elementary events in E of a certain information source are fixed by the experimental set up $< E, \mathcal{F}, P >$. Yet, in section (2.1) we already pointed out that Shannon's formalism basically doesn't restrict as what counts as an information source. On these grounds, we can 'manipulate' the elementary events by stipulating the source. For instance, in our urn-example, the drawing of a red or blue ball each counted as an elementary event. By stipulating a slightly adjusted source–let's say we additionally consider a clock which tells us the time at which a ball is drawn to be part of the source–we might get elementary events like 'red ball drawn at 12:37am' or 'blue ball drawn at

midnight'. Our (arbitrary) adjustment of the source then in turn might yield a completely different probability distribution $P_S(X)$.

Changing the probability distribution of the information source by partitioning, means to change Quantity Information. Stipulating (an adjusted) information source, might change the respective probability distribution, hence also changing Quantity Information. Quantity Information can then only in so far be regarded an 'objective' or intersubjective information measure, as the experimental set up and the experiment are somehow a priori fixed. However, in the classical case we're not only able to stipulate myriads of experimental set ups, but also experiments. Since a natural fixation is missing, it appears to be almost entirely conventional how either of them should be a priori fixed.

How does this finding influence our view of Type Information? Changing $P_S(X)$ of the source affects Type Information to the extent that now a different set of (type) sequences is deemed to be producible by the source. Hence, Type Information doesn't remain unaffected and seem to be completely governed by the contingencies of the users who set up the communication system. Note though, that the above considerations don't have any bearing on the abstract nature of Quantity- and Type Information!

The Fidelity function

Reconsidering our definition of information, '*to be counted as a success*' may import another conventional element in Shannon's formalism. For instance, how could we possibly know what's encoded in a sequence of 1s and 0s without a prearranged, purely conventional agreement? What determines if a transmission was successful? As Duwell [Duwell, 2008] pointed out, Shannon's original paper indeed doesn't clarify a success criterion. The only clue we're offered in that regard, is the so called fidelity function (2.23) we've already encountered in Chapter 2.[14] As the name suggests, the function ought to measure the fidelity, i.e. measuring the rate of (or accuracy of) reproduction of a message.

[14]Notice that Shannon's original paper isn't simply used as a straw man here. As far as the author is concerned, the subsequent information-theory-literature remained largely silent on defining a success criterion too.

Solely based on the fidelity function, Duwell could nevertheless identify three success criteria (unfortunately, without giving them clear names) coming in different degrees of the same idea:

1. A one-on-one mapping of physical features of the tokens. This view proposes that the alphabet of the source has to be exactly reproduced at the destination. In an everyday life communication scenario, the sound spoken into a telephone A may be converted into an electric signal and then at the destination (here telephone B) be reproduced into sound again.[15]

2. A mapping which restricts certain physical characteristics. This notion is similar to the above, except that we're less strict with the term *exactly*. So we might, *based on convention*, allow for the reproduction of different types of tokens at the source. Let's suppose, we want to send an email with the content: '$HELLO$'. According to the former success criterion, only recreating an exact copy of our email would count as success. In our case though, we might allow for the reproduction of '$\mathcal{HELLO}$' at the source (but not '$\mathcal{HEXXO}$') . Even though, the physical characteristics of what's sent from the source and what's received at the destination are different in both cases (i.e., the letters have different shapes), we might call a transmission successful only in certain cases. Hence, the success criterion restricts which tokens of the type produced by the source are admissible.

3. An arbitrarily chosen one-on-one mapping function *not* based on any physical characteristics at all. On this view, we might treat information sources as producing an abstract sequence purely based on the Shannon formalism, allowing a success condition dependent on an arbitrary one-on-one function between source and destination. Such a success criterion might work without considering any

[15]Note that the example is an *idealized* everyday scenario, since a real setting is (to some extent) always obscured by some external influences. The so created noise hampers an exact recreation at the source, making it necessary to allow for a certain margin of error. Usually, a distance measure that compares the tokens produced by the information source and the tokens reproduced at the destination is used.

physical characteristics that the source and destination share at all. What counts, is a matter of convention.

Note that none of the success criteria above, acts as a 'natural' success criterion. The functioning of Shannon's theory doesn't specify a preferred success criterion so that choosing such, is completely conventional. According to the third success criterion for instance, we're then allowed to stipulate a communication system in which in comparison to the source, totally different sequences of tokens can be reproduced.

Choosing the experimental set up, deciding which experiment to perform and determining what counts as a success, imports substantial conventional elements into Shannon Information in the classical case. Note though, that the import of conventional aspects neither changed our view about the abstract nature of Quantity- nor Type information.

5.6 Conclusion

Our analysis in the previous sections has shown us that Shannon Information can be interpreted in more than one way. We argued that the interpretation of compressibility in context of communication and the minimum average code word length are tightly interwoven. On the other hand, we demonstrated that $H(X)$ can also be interpreted as *one among many* measures of uncertainty. However, the latter notion is at odds with Shannon's Noiseless Coding Theorem, which states the existence of a *unique* compressibility scheme. Thus, for the description of any kind of communication systems (or likewise scenarios in nature), we only regard the former non-uncertainty based notion.

Examining the Shannon formalism under these conditions, we first argued that the semantic/syntactic distinction of information (introduced in the Chapter 1) indeed holds. The Shannon formalism allows to characterize different aspects of communication systems and doesn't adhere to any semantic properties. Thereafter, we opted for a further distinction of Shannon's syntactic information, namely Quantity- and Type Information. Though for different reasons, both notions are abstract in respect

to their ontological status; Quantity Information, because a calculation based quantity is always something abstract and Type Information since we identified it (based on a type/token distinction) as a universal. On our view, Type Information can be encoded with concrete tokens, but mustn't be confused with them. The tokens only function as a mean to solve the 'problem of communication' and *successfully* reproduce the message at the destination.

Analyzing the extent of conventionality featured in Shannon's theory showed that those who set up the communication system have the freedom to determine what qualifies as successful communication. Furthermore, there are no constraints which restrict what counts as a natural information source and its respective outcome; it's almost entirely conventional. In the classical case, Shannon Information may then only be useful once these conventional elements are somehow fixed.

Regarding the ontological status of Shannon Information in the classical case, we can conclude that $H(X)$ is a largely conventional and abstract entity, independent of any notions of semantic Information.

Chapter 6

Interpreting Kolmogorov Complexity in the Classical Case

"Our definition of the quantity of information has the advantage that it refers to individual objects and not to objects treated as members of a set of objects with a probability distribution given on it."
[Li and Vitanyi, 2008, p. 603]
– *Kolmogorov*

In the previous chapter we dealt with Shannon's combinatorial approach which merely focuses on an *ensemble* of typical messages. Algorithmic information in contrast, only pays attention to *particular* messages or objects. At first, one could get the impression that the algorithmic approach is indeed concerned with a concrete entity. Is information, given in this context, something concrete after all and hence at odds with the abstract notion of Shannon Information? To answer this question, let's turn to a rigorous interpretation of the algorithmic information measure (i.e. Kolmogorov Complexity) and examine to what extent its ontological status is different from that of Shannon Information.

6.1 Uncomputability, Unpredictability and Uncertainty

In comparison to Shannon's information measure, the algorithmic approach to measuring information, suffers much less from the deficits of ambiguous terminology. As we've seen earlier, the definition (3.8) tells us when a sequence is simple or not. However, as mentioned in section (3.3.2), we generally can't compute $K(x)$. As a consequence, we can never *know* if we've indeed found the shortest algorithm or whether a sequence is random. So, in the context of the uncomputability of $K(x)$ we are then confronted with *unpredictability*. One may then equate 'unpredictability' with 'uncertainty in prediction' and conjecture that $K(x)$ is a measure of uncertainty. As we'll show in the following though, unpredictability here refers to our epistemic limitations *about* $K(x)$, i.e. giving a prediction about the length of the shortest program that describes x. Whether $K(x)$ itself is about uncertainty, i.e. if Kolmogorov Complexity is a measure of uncertainty in prediction, is a different matter.

Uncertainty *about* K(x)

As Chaitin points out [Chaitin, 1975], [Chaitin, 1982], [Chaitin, 1986], the limitation to compute $K(x)$ is not a flaw in the definition, but a consequence of Kurt Gödel's (1931) closely related *Incompleteness theorem.* Gödel's proof is based on paradoxes as "This statement is false" or the Berry paradox we've seen before (see end of section (3.1)). By rewriting the Berry paradox in terms made suitable for a computer program,[1] Chaitin shows that in a formal system of complexity n, it is impossible to prove that a particular binary string is of complexity $> n + c$ (where c is a constant independent of the regarded system).

An example for clarification. We assume to be confronted with the task to determine the Kolmogorov Complexity of a very long binary sequence $x_{example}$ with a given reference $\mathcal{TM}$. So why is it, that we can't

[1] Rewritten the paradox might look like "Find a series of binary digits that can be proved to be of a complexity greater than the number of bits in this program." [Chaitin, 1975]

compute $K(x_{example})$ and predict whether that sequence is random? From our definition of randomness it follows that in general any shortest program p (i.e. in fact $K(x)$) is necessarily random, for if it was not, we could find a shorter program p^* that in turn generates p, which in turn generates x. From our reflections in section (3.2) on the other hand, we know that any $\mathcal{TM}$ can be 'simulated' through the input on a $\mathcal{UTM}$. This input has a certain complexity and can, according to Chaitin, be conceived as a formal system of complexity n, such that it can't prove the complexity of strings larger than itself $> n + c$. If we then want to determine whether the sequence $x_{example}$ is random, where the complexity of the sequence is $> n + c$, our reference $\mathcal{TM}$ of complexity n won't halt. If our $\mathcal{TM}$ were indeed to halt, it then either would have algorithmically produced a sequence larger than itself (meaning that $x_{example}$ is in fact non random) or proven the complexity of a sequence larger than itself. So for proving the randomness of a given sequence, the complexity of the reference $\mathcal{TM}$ always has to be larger than that of the given sequence itself. This implies, because $K(x)$ is random, that we require a reference $\mathcal{TM}$ of greater complexity for proving that a program is truly a minimal one for a particular sequence x.

What the example has then shown us, is that we are uncertain *about* the value of $K(x)$, not whether $K(x)$ denotes the upredictability or uncertainty of some event. Once we have the necessary resources and our reference $\mathcal{TM}$ halts, providing us with a value of $K(x)$, we know to have found the shortest program p which uniquely describes x–there is not any kind of uncertainty left *about* $K(x)$.

K(x) - Not a Measure of Uncertainty

Let's for sake of completeness quickly examine the relation of measures of uncertainty and $K(x)$. The question if $K(x)$ is a measure of uncertainty (in prediction), might appear as an unusual one; denoting the shortest length of an algorithm that generates a sequence doesn't appear suitable as a characterization of uncertainty. One of the motivations for the development of algorithmic information was to escape the probability based notion of Shannon Information. At first sight, probabilities–the very ba-

sis to measure uncertainty–aren't featured in the definition of $K(x)$ (3.5) at all.

However, when we introduced the notion of universal probability $m(x)$ in Chapter 4, we saw that Kolmogorov Complexity could equally well be defined as $-\log_2 m(x) = K(x)$ (4.11). Does $K(x)$ qualify as a measure of uncertainty after all?

In order to classify as a measure of uncertainty, Uffink pointed out a number of criteria (see section (5.1)), which $K(x)$ fails to agree with. Remember, the basic of idea of a measure of uncertainty is to measuring the concentration of a probability *distribution*. Regarding our definitions of Kolmogorov Complexity, no such probability distribution is given in the first place; $K(x)$ is either defined as the shortest *length* of the program that uniquely generates x or as the negative logarithm of the universal probability $m(x)$ (which denotes the probability to obtain the binary program p by flipping a coin, for instance; compare section (4.2)). While the former definition quite obviously offers no grounds to speak about the concentration of a probability distribution, the latter doesn't either. We mustn't confuse the universal probability $m(x)$ with a probability *distribution*![2] Uffink's approach to determine uncertainty in prediction can only be applied to probability distributions, not to single probabilities. Hence, Kolmogorov Complexity is not a measure of uncertainty $U_r(P, \mu)$.

6.2 Hidden Semantics?

We now analyze to what extent algorithmic information suffers from semantic backlashes, by examining the relation of random and meaningful sequences. Arguably, one could construct the argument that non-random sequences (sequences with patterns) have to entail meaning. And in fact, regarding meaningful sequences seems to be one of the reasons to develop the notion of Kolmogorov Complexity in the first place. For instance, remember the epigraph of Chapter 4 about Tolstoy's *War and Peace*, in

[2] In our context, one might regard $\sum_x m(x)$ as such a 'probability distribution'. Since $\sum_x m(x) \leq 1$ denotes only a semi measure, one possibly has to make further adaptations to analyze $m(x)$ with Uffink's formalism at all.

which Kolmogorov pondered how to determine a probability distribution for such a 'sequence'; should the entire novel be included in a set of all possible novels and then merely be treated like any other $\varepsilon -$ typical sequence in Shannon's formalism?

Clearly, finding such a probability distribution is 'challenging' (to say the least) and while it might be true that every meaningful sequence $x_{meaning}$ displays certain patterns,[3] we mustn't conclude that the formalism of Kolmogorov Complexity assigns a value of semantic information to any object or sequence x. Whether a sequence x is meaningful or not doesn't matter; without any extra assumptions, nothing in the definition of $K(x)$ indicates a relation with semantics. $K(x)$ merely denotes the length of an algorithm that uniquely generates x as the output of a $\mathcal{TM}$. Purely based on *syntax*–the structure of the symbols forming patterns and redundancies–a given message might be compressed such that an algorithm shorter than the message itself can be found that can produce it. Considering the arrangements of (code-)words and their relation in a sequence is purely syntactic, not semantic!

6.3 Repeating the two way strategy?

Key for the understanding of the ontological status of $K(x)$, we'd like to know whether Algorithmic Information is something 'physically concrete' or something abstract. At first, one could conjecture that by avoiding the probability based ensemble view of Shannon Information and by instead only regarding individual messages or objects, algorithmic information is indeed something concrete. But then the term *algorithmic* suggests that we actually deal with algorithms–an *abstract* set of rules determining a specific problem-solving activity. For avoiding the tension between particular concrete objects and non-concrete computer programs, we can

[3] In order to be meaningful, a sequence like $x_{meaningful}$ arguably has to be formulated in a certain language. Languages have to follow certain rules and, as we've learned earlier, they show patterns according to the probability distribution of their respective alphabet. Additionally, there are 'higher order patterns' like, certain letters often follow other letters (like 'q' and 'u') or certain words usually don't follow certain words ('the' and 'the', e.g.). Instead of the assumption that every outcome of our information source is independent of the previous one, one may regard Markov processes instead.

repeat the 'two way strategy' from the previous chapter by yet again distinguishing between Quantity- and Type information.

Algorithmic Quantity Information

Just like in the case of Shannon Information, we can easily point out that Quantity Information–this time the $K(x)$-version–merely quantifies algorithmic information. The value of $K(x)$ denotes the shortest length of an algorithm generating a sequence x and nothing in this formalism bears the characteristics of concrete objects. Even considering the definition based on the universal distribution $m(x)$ (4.11) won't change our view. Remember, $m(x)$ denotes the probability that a hypothetical random process (with binary outcome) generates a program p which in turn uniquely generates a sequence x on our chosen reference $\mathcal{TM}$.

Quite analogue to the previous chapter, we may then describe Kolmogorov-Complexity-based Quantity Information as

> *Quantity Information is a function dependent on the redundancy (patterns) of a sequence x, which neither depends nor quantifies which kind of tokens x is made of.*

Once again, we can conclude that Quantity Information is an abstractum and merely suffices to quantify algorithmic information.

Algorithmic Type Information

The question of '*what is actually quantified?*' is a much more intriguing one and it's far from obvious that in the context of Kolmogorov Complexity the term *Type* Information is appropriate. As stated in the epigraph by Kolmogorov (see the beginning of this chapter), $K(x)$ refers to individual 'objects' and not to members of a set determined by a probability distribution. If we perceive these individual objects as *concrete particulars* does that imply that algorithmic information is something concrete?

Let's first take a look at the 'individual objects'. Certainly we can spatio-temporally localize many of these objects; perhaps we regard the redundancies of the atomic structure of a salt crystal or consider the characteristics of a painting. These kind of objects are certainly concrete.

However, note that the term 'object' has to be understood loosely here, in the sense that we could also look at the patterns of things that are usually not considered objects. So we could e.g., as well use the formalism of $K(x)$ to examine certain patterns of the weather, sun eclipses or the ordering of books in a shelf. While entities like 'the weather' or 'a sun eclipse' might be difficult to conceive as an object, *the patterns* displayed by these phenomena still bear on concreta. Basically, the formalism of $K(x)$ doesn't come up with any kind of restrictions of usage at all. In principle, we can examine whatever observable 'object' we like, as long as some description method ensures to extract a binary sequence x from the characteristics of that 'object' (more shall be said about such a procedure in the next section). So while observable objects (or rather patterns) are based on concrete particulars, we mustn't commit the fallacy that algorithmic information is concrete, too.

Remembering our insights from Coding Theory and repeating the argument based on the type/token distinction of the previous chapter, we may argue to find the same *algorithmic type information* to be *instantiated* in the characteristics of many different 'objects'. Think, for instance, about the *golden ratio* based on the sequence of the *Fibonacci numbers* $\{0, 1, 1, 2, 3, 5, 8, 13, 21, 34, ...\}$.[4] Examples in nature where the golden ratio can be found range from 'objects' like DNA molecules, snail shells, fruits and vegetables (e.g, spiraling patterns in pineapples), hurricanes up to spiral galaxies. In all these instances, certain characteristics of the 'objects' act as means to encode a certain universal sequence–an algorithmic information type sequence.

Lastly, we can then describe Algorithmic Type Information as

> *Algorithmic Information is what is instantiated by the patterns of some kind of objects or tokens, such that the description of these patters can be uniquely generated by an algorithm.*

Notice that this definition is (intentionally) of utmost generality and not restricted to any special class of objects or tokens. According to our def-

[4]The golden ratio is the limit of the ratios of successive terms of the Fibonacci sequence.

inition, Algorithmic Type Information is *instantiated* by the characteristics/patterns of objects. On that view, the term *type* is appropriate, as algorithmic information appears to be an *abstract sequence* which can be multiply instantiated by the means of concrete tokens.

6.4 To what extent is Algorithmic Information conventional?

Similarly to section (5.5) in the previous chapter, we analyze to what extent Kolmogorov Complexity bears conventional elements. In the following, we begin by looking at the definition of randomness, the role of choosing a specific reference machine and a certain program language. Thereafter, we turn our attention to coarse graining and how to describe objects. As we'll see, it turns out that especially the latter notions have enormous bearing on the conventionality of Algorithmic Information.

6.4.1 The Definition of Randomness

First a few words about the 'exact definition of randomness' in the context of Kolmogorov Complexity. In the introductory chapter of Kolmogorov Complexity, we have learned that the majority of sequences is in fact random. However, so far we haven't quantified a precise threshold for when a sequence clearly counts as 'random', to actually jump to such a conclusion. Chaitin, e.g., argues that

> "The exact value of complexity below which a series is no longer considered random remains somewhat arbitrary." [Chaitin, 1975]

Choosing a certain value in order to determine the randomness of a sequence then may contain a degree of conventionality. So we could, for instance, set the threshold for a sequence of length N to $\frac{N}{2}$; every series for which an algorithm of half its length (or below) can be found, might then not be considered random. Yet, whether or not a given sequence is indeed to a certain extent compressible, certainly doesn't change our view

about the ontological status of information. What's at stake in this subsection is whether the definition of randomness is conventional, not that of Kolmogorov Complexity. For the sake of this thesis, we thus shouldn't be bothered by the exact definition of randomness any further.

6.4.2 Universality & the Invariance Theorem

In Chapter 3, introducing the main formalism of algorithmic information, we encountered the so called *Universality* and *Invariance Theorem* as features of Kolmogorov Complexity. Both these features are concerned with making choices that are crucial for the calculation of $K(x)$–choosing a reference machine and a program language. To what extent are these choices a matter of convention?

Choosing a reference machine

In the context of Kolmogorov Complexity, *Universality* is expressed as $K_{\mathcal{U}}(x) \leq K_{\mathcal{A}}(x) + c_{\mathcal{A}}$ (3.7), describing that choosing a different reference machine (denoted by the indices '$\mathcal{U}$' and '$\mathcal{A}$'), in order to calculate the shortest program p that calculates x, merely yields an additional constant $c_{\mathcal{A}}$. That is because any $\mathcal{TM}$ can in principle be 'simulated' by another $\mathcal{TM}$, where $c_{\mathcal{A}}$ is the length of the program needed for such a simulation.

Admittedly, choosing a certain kind of $\mathcal{UTM}$ is then still purely arbitrary and a choice of convention, but that doesn't matter. Save for a constant, the value of Kolmogorov Complexity is *universal* and hence independent of any kind of reference machine. Key for the value of $K(x)$ are the characteristics of x (for instance, if the sequence x shows 'compressible' patterns), not the machine which it is calculated on.

It's insightful to imagine a situation in which the contrary was the case, i.e. Kolmgorov Complexity was dependent on a certain reference machine. In that case incorporating $\mathcal{TM}$s as an auxiliary tool wouldn't have been a clever move, because the initial intention of algorithmic information to measuring the information content of individual objects x would now also depend on something other than x. Constructing such an information measure would be completely arbitrary and hardly tenable

for any scientific purpose. However, the very reason we use $\mathcal{UTM}$s in the first place, is it that we can benefit from the fact that they can simulate each other. Using $\mathcal{TM}$s only acts as a mean to make the notion of shortest algorithm p unambigious.

Choosing a language

Besides picking a reference machine, we also have to pick a certain program language like Java, Lisp, etc., which can be conceived as an encoding scheme. Once again, we're seemingly faced with a conventional choice–which program language or encoding scheme shall we choose?

However, similarly to the case of Universality in the former subsection, the choice of a language certainly is conventional, but according to the *Invariance Theorem* $K_1(x) \leq K_2(x) + O(1)$ (3.6), the choice doesn't affect the value of $K(x)$ (save for a constant). Effectively, conceiving a program language as a code, the shortest program p can be *realized in multiple ways* (i.e. in multiple program languages). Once again, what determines the value of $K(x)$ are the characteristics of x itself, not the 'external' choice of a program language.

6.4.3 Coarse Graining & Description

In the case of Shannon Information, the experimental set up, the experiment and the choice of a success criterion are contingent on the interests of those who set up the communication system. Can we find similar conventional elements in the formalism of Kolmogorov Complexity?

So far, we only regarded instances in which the (shortest) algorithm p *exactly* describes the sequence x; all these instances have a fidelity of one. That's because by using the notion of prefix-free codes, we assured that every program p is capable of uniquely generating *only one* specific sequence x, on a deterministically operating $\mathcal{UTM}$. So unless a non-deterministic reference machine is used (for instance, a probabilistic $\mathcal{TM}$), a one-on-one mapping

$$p_i \rightarrow x_i, \tag{6.1}$$

with $i \in N$, for all possible N descriptions, from a particular *shortest* program to a specific sequence is ensured.[5] Thus, whenever for instance, a sequence

$$x_1 = 10101010101010101010101010101010,$$

has slightly changed (notice, the last digit has switched its value)

$$x_{1'} = 10101010101010101010101010101011,$$

we don't have to rely on the very same program p_1 (that originally generated x_1), to obtain the modified sequence $x_{1'}$. Instead, we can try to find the shortest program $p_{1'}$ that precisely gives out $x_{1'}$. In the context of Kolmogorov Complexity, there's no need to evoke a fidelity function or any conventional success criteria like in the context of communication. However, when considering that the above sequences are *descriptions of* something (perhaps an object like a painting or a weather phenomenon), we start to run into two major elements of convention quite similar those ones the in the case of $H(X)$ (choosing experimental set up and deciding on a partition).

Coarse Graining

Let's start with an example and assume that various scientists want to characterize if the weather shows certain patters. Based on a month long observation, a meteorologist writes down a '1' for a rainy day and a'0' for a day without any precipitation, obtaining e.g., sequence x_1. However, are there any non-conventional reasons for why to choose a such specific 'resolution'? Why does a meteorologist choose 'days' as the preferred time intervals; why not intervals of 17 hours or of 3 and a half days? For instance, a climate scientist on the other hand, might rather not be interested in the observation data of single days, but instead of long-term observations over decades or centuries. Clearly, we expect the complexities of 'objects' (here the weather), to change when a different 'resolution'

[5] Not considering only the shortest programs yields a surjection (instead of a bijection like in the case of one-on-one mapping), i.e. for every sequence x_i there'll be multiple programs to generate that sequence.

is chosen.

Regarding the problem of resolution, statistical mechanics and thermodynamics are faced with similar discussions when it comes to *coarse graining* (see e.g. [Denbigh and Denbigh, 1985, §3]). In short, coarse graining is the rescaling of a phenomenon into cells or units close to the uncertainty of our measurement (or our interest). In the context of statistical mechanics, the cells of the phase space are coarse grained (compare the example of algorithmic entropy in section (4.5)), so it is often argued, according to human choice. Under the circumstances specific to the context of statistical mechanics, i.e. trying to account for irreversibility for instance, the question whether coarse graining is 'subjective' remains disputed.

What can we take over from this debate to our current case? Statistical mechanics ought to explain the purely phenomenological 'laws of Thermodynamics'. As such, in converse Thermodynamics offers us constraints through empirical observations and provides statistical mechanics a context to which coarse graining can be applied (the phase space of a system).

The bare formalism of Kolmogorov Complexity on the other hand, doesn't come equipped with any constraints or a particular context it's supposed to be applied to. As pointed out in the previous section, the scope of the formalism is virtually applicable to anything. Such a generality doesn't offer any constraints; this means, that unless we're offered some kind of boundaries as in the case of statistical mechanics and Thermodynamics, that the coarse graining in the case of $K(x)$ is completely contingent on the user.

Returning to the example of our scientists, some may want to examine the regularities of single raindrops, others the mean rainfall over the last millenia. Depending on the chosen resolution, i.e. coarse graining, objects can virtually have infinitely large complexity. One may then find very different Algorithmic Quantity Information values and different Type Information sequences depending on the completely conventional stipulated coarse graining of the user.

Kolmogorov himself seemed to be aware of these kind of troubles and

suggested to prefer the notion of mutual complexity $K(x : y)$ (3.11), instead of the complexity of $K(x)$ stating that

> "Actually, it is most fruitful to discuss the quantity of information "conveyed by an object" (x) "about an object" (y) [...] The real objects that we study are very (infinitely) complex, but the relationships between two separate objects diminish *as the schemes used to describe them become simpler* [own italics]. While a map yields a considerable amount of information about a region of the earth's surface, the microstructure of the paper and the ink in the paper have no relation to the microstructure of the area shown on the map." [Kolmogorov, 1965, p. 6]

In the quote e.g., the map represents the object x which conveys something about object y (the surface of the earth). However, it's not clear which coarse-graining procedure ought to determine such 'simpler schemes' in order to describe the relation between the objects x and y.[6] Shifting the attention from a single object x to the relation between object x and y, merely shifts the problem of coarse graining from single objects to the relation of objects. The entirely conventional aspects of coarse graining remain the same for $K(x)$ and $K(x : y)$ though.

Extracting Patterns–A Problem of Description

Another, not less important point of convention comes into play when we ask how we obtained sequence x from an outside state of the world; we've to know how to describe the patterns of objects (which we suspect to instantiate Algorithmic Type Information) with ones and zeros. In other words, we're looking for a description method that extracts the patterns of objects. But who or what determines which patterns to extract–what are the sequences $x_1$and $x_{1'}$ (or in general, every kind of sequence) actually descriptions of?

[6] Even though maps usually come equipped with a scale (a ratio of a distance on the map to the corresponding distance on the earth's surface), such that the coarse graining is fixed, the choice of the scale is completely conventional.

First we should note that we mustn't commit the fallacy to think the variable x in $K(x)$ objectively represents an 'object x'; instead x denotes a sequence that *merely describes* that object. A description of something is entirely different than that 'something' itself.[7] So without a natural description method specifying how to get from patterns of objects to a sequence x describing these, we can't apply Kolmogorov Complexity to more than just somehow already given binary sequences. As remarked in [Gruenwald and Vitanyi, 2008, p. 315], one may conclude that[8]

> "algorithmic information misses "aboutness" (sic), and is therefore not really information."

If we want to denote the complexity of objects, the problem runs down to the absence of a 'natural description method'. Since it's extremely puzzling how a single mechanism could possibly account for the extraction of certain patterns in a myriad of different kinds of objects, we instead have to rely on a conventional method. Recall, in our examples about the meteorologist and the climate scientist we simply imposed completely conventional description methods (e.g.,'1' for rain, '0' for no rain). Just because of some kind of conventions–solely considering if it was a rainy day, not a cold one, for instance–the scientists were able to extract patterns from the weather in the first place. Thus only after it's agreed on how to describe certain patterns, it might be useful to perceive objects as instantiating Algorithmic Type Information. Without further conventional contextualization, $K(x)$ may only provide an information measure of an a priori given binary string x, independent of any object or outside state of the world.

6.5 Relation between H(X) and K(x)

For the framework of this thesis it's a crucial aspect to point out the relation between different notions of syntactic information. Let's begin to

[7] N.B. that Kolmogorov in the quote above as well as in the epigraph, seemingly fails to recognize the distinction between 'object' and 'sequence'.

[8] The statement emerged in context of a workshop, with many participants being contributors to the handbook in which the reference was published in.

compare our so far made results of Shannon Information and Kolmogorov Complexity in the classical case.

The Semantic/Syntactic Distinction

In the introduction (Chapter 1), we claimed that the term information refers to at least two different concepts, semantic information and syntactic information. As we have seen in section (5.3) and (6.2), neither Shannon Information nor Kolmogorov Complexity bear any formal relation to semantic information. Nevertheless, for special purposes one may link either of the formalisms–in one way or another–to semantic information. For instance, when we think of communication in an every day sense, we usually want to convey something meaningful, i.e. a form of semantic information. It's not prohibited to use Shannon's formalism to formally describe a communication system in which such every day communication takes place. In the same vein, one may analyze meaningful texts and their patterns with the help of Kolmogorov Complexity.

Various interpretations

At first, we encountered at least two different interpretations of Shannon Information–$H(X)$ as a measure of uncertainty and $H(X)$ as the optimal statistical compression rate of messages emitted by an information source. Based on the groundwork of [Uffink, 1991] and [Timpson, 2013], we demonstrated (in a lengthy but necessary treatment) that only the latter notion of $H(X)$ is adequate for our analysis. After all, we continued to just examine the communication based interpretation of $H(X)$.

Kolmogorov Complexity on the contrary, only offers one interpretation from the beginning. $K(x)$ denotes the length of the shortest program p that generates the sequence x. Even though one might regard the length of such a program as dependent on the universal probability $m(x)$, we aren't faced with an (entirely) new interpretation of $K(x)$ after all. The notion of universal probability merely establishes a connection between the length of programs and probabilities; based on the Kraft inequality (4.6) one can deduce the probability $m(x)$ that a random process might

exactly generate the algorithm p that generates the sequence x.

Physical domain

Let's start comparing the physical domain, i.e. those features in nature each of the information measures can be applied to. While Shannon Information is restricted to communication systems, the notion of Kolmogorov Complextiy can in principle be used to describe any kind of object.

Regarding Shannon Information, note that 'communication systems' are not as restrictive as they might first appear to be. Shannon's theory is neither limited to the actual scenario of Alice and Bob communicating, nor to a 'natural' or any other specific kind of communication system. As pointed out in section (2.1), there are only a few constraints as what may count as parts of a communication system. Basically, all that's required is a source with a probability distribution over some kind of output states that have to correlate (according to a completely conventional success criterion) to some states at the destination.[9] Since 'source' and 'destination' can be understood with utmost generality, the scope of Shannon's theory is almost endless.

As argued above, the formalism of Kolmogorov Complexity doesn't come equipped with any kind of boundaries to what it can be applied to–given a description method, essentially all kinds of 'objects' may be analyzed in regards of Algorithmic Information.Whereas $H(X)$ has to be applied to the wider sense of communication systems, $K(x)$ is even less restricted, only having to look at whatever kind of object, with no need to adhere to probability distributions. Note in addition, that like in the case of Shannon Information, the term 'object' should be understood with utmost generality, such that the term may also refer to an order of things or systems like hurricanes or the solar system.

Overall, both the notions of $H(X)$ and $K(x)$ can be applied to a seemingly endless physical domain.

[9]Note that the measurement apparatus/experimenter can be regarded as destination too.

The Type/Token Distinction

As we have seen, Timpson's 'two way strategy' [Timpson, 2004] to differentiate between Quantity- and Type Information in the case of Shannon Information, could also be successfully applied to $K(x)$. Based on an argument of scientific representation, we identified that Quantity Information merely quantifies Type Information. While Shannon- and Kolmogorov Quantity Information are abstracta for the reason that quantities can't be spatio-temporally located, we had to develop more elaborate arguments for the respective notions of Type Information.

One explanation why we could identify the abstract notion of Type Information featured in Shannon Information as well as in Kolmogorov Complexity lies in an *overarching theme*. Our starting point for the comparison of Shannon information and Kolmogorov Complexity in Chapter 4 was the introduction of Coding theory. As has become evident throughout this thesis, both Shannon information and Kolmogorov Complexity crucially depend on the insights of coding. Whereas Shannon Information denotes the optimal rate for the minimum *average* codeword length for a pool of messages, Kolmogorov Complexity determines the minimal 'codeword length' of an *individual* message. Both information measures rely on the insights of Coding Theory and represent different strategies to optimize coding–a probability based and an individually based approach. At the end of Chapter 4, we even showed that asymptotically, expected Kolmogorov Complexity equals Shannon Information $\sum p(x)K(X) = H(X)$ (4.16).

The reason why we can conclude that Type Information is an abstract notion in both cases, is that our here regarded syntactic information measures are bridged by Coding Theory. The multi-realizability of messages allows us to successfully apply the type/token distinction to both cases! Messages are an abstract notion encoded by the means of tokens. In the case of Shannon Information we concluded that virtually all kind of tokens (emitted from a source) suffice to encode messages. In the case of Kolmogorov Complexity, we argued that virtually all kinds of 'objects' suffice to encode messages.

Conventionality

The classical-case analysis revealed that both information measures are entirely user dependent in the following sense: In the case of $H(X)$ the set up of the experiment (the communication system) is completely up to the user. Additionally, the 'choice of the experiment' requires a certain partition, i.e. we may partition the outcome of the source according to our interests. Lastly, because of the lack of a natural success criterion, the user has to evoke some conventional success criterion. Altogether, both Shannon Quantity Information and Shannon Type Information are then merely stipulated by human choice.[10] In a similar fashion, Algorithmic Quantity Information and Algorithmic Type Information almost entirely depend on the description method and the subsequently stipulated coarse graining, conventionally chosen by the users.

Save for the success criterion,[11] the conventional elements in Shannon Information and Kolmogorov Complexity show in fact great resemblance. In the case of $H(X)$, the set up of the experiments matches the choice of a description method we encountered in context of $K(x)$. Each procedure ensures to pick out the tokens we might apply the respective formalisms to; the elementary events (the outcomes of the source) in context of communication systems and the kinds of patterns we extract from objects when considering the complexity of such. In addition, we can then for each information measure choose a 'resolution of the tokens'. Regarding Shannon Information, we have to choose a certain partition in order to decide based on which characterstics of the tokens we want to distinguish them (i.e.,'perform our experiment' on). For instance, do we differentiate between different fonts and styles of the letter 'A' (5.8) as possible outcomes of our source, or do they all count as instances of 'A'?. Quite similar, we have to decide on our coarse graining in the case of $K(x)$. Considering our weather-example, one might agree to describe (the patterns of) rain and no rain with '1' and '0'. However, without a fixed coarse grain-

[10] To be clear, such a stipulation won't change either notion of Type Information to be a concrete entity though.

[11] Remember a success criterion isn't required in the case of $K(x)$, since (on a deterministic $\mathcal{TM}$) 'successful' is guaranteed by the usage of prefix free codes (i.e. programs).

ing–here stipulating a certain time interval (perhaps minutes, hours or days)–it won't be possible to unambiguously extract a string of ones and zeros from the weather. As upshot of this, we conclude that for quasi similar reasons, Shannon Information and Kolmogorov Complexity are highly conventional.

A few remarks. Of course the import of conventions doesn't imply that the here regarded syntactic information measures are completely useless. The formalism of Shannon Information is 'objective' (in the sense that it's intersubjective) once the set up of the experiment, the experiment and the success criteria are agreed on, such that the probability distribution of the outcomes of the information source is fixed. For practical application we can often agree on these criteria. In a classical communication scenario with two English speakers e.g., we can calculate the channel capacity based on the probability distribution over the letters. Without Shannon's formalism, much of modern telecommunications wouldn't be successful. In the same vein, we can use $K(x)$ once we fixed the the issues concerning how to extract patterns and describe them. The notion of Kolmogorov Complexity helps us for instance, to compress otherwise large data files.

6.6 Conclusion

Kolmogorov Complexity

Our analysis of Kolmogorov Complexity has shown that we might be never able to predict the exact value of $K(x)$ for a certain sequence x. Such an unpredictability is not a defect of the formalism, but a result of the Halting Problem of $\mathcal{TM}$s. Motivated by the discussions about uncertainty in the case of Shannon Information, we argued that $K(x)$ isn't a measure of uncertainty. Thereafter, we could demonstrate that Algorithmic Information is indeed solely a measure of syntactic information and doesn't bear on notions of semantic information. We could then demonstrate that the notions between Quantity- and Type Information of Chapter 5 could be translated to the current case. For different reasons, both

notions are abstract; the former since quantities are always abstract, the later based on a type/token distinction. Eventually we illuminated that Algorithmic Information lacks a natural 'description method' and thus bears on conventional elements stipulated by the user. Additionally, further conventional elements are imported by the degree of coarse graining.

We conclude that Algorithmic Information is an abstract and largely conventional entity, which is independent from any notions of semantic information.

The relation of $H(X)$ and $K(x)$

Comparing $H(X)$ and $K(x)$ showed strong similarities between both information measures. As claimed in the introduction, neither Shannon Information nor Kolmogorov Complexity bear any relation with semantic information. Both formalisms eventually allowed us to pick out one unique interpretation. As we've seen, each of these interpretations can be applied to a virtually endless amount of communication systems or objects respectively. The main insight from the framework (comparing $H(X)$ and $K(x)$) of this thesis thus far, is that both information measures are linked through Coding Theory and are in one way or another concerned with encoding. Ultimately, we could come to the same result regarding Quantity- and Type Information–all these notions of syntactic information are abstract. Abstract messages or sequences can be generalized by a great variety of tokens. However, in both cases such a generality comes with the price of not having natural constraints; no natural experimental set up, experiment, and success criterion in the case of Shannon's theory and (quite similar) no natural description method and coarse graining that extract the patterns of objects. As a result, both information measures are in the classical case at the bottom highly conventional. In both cases, the absence of natural constraints has to be fixed by the users of the theory.

Chapter 7

Quantum Information theory

"Many classical quantities, e.g the Shannon entropy, have been successfully generalized to quantum information and have become useful and powerful tools to understand and further develop quantum information theory. In the case of Kolmogorov complexity, though, the way to do so is not straightforward."[Mora et al., 2006, p.2]

– Mora et al.

QUANTUM mechanics is regarded as our most successful scientific theory. It describes systems at what is thought to be their most fundamental level. Since the point when incorporating information theoretic aspects gained momentum in physics (beginning of the early 1990s, see Chapter 1), the 'quantum world' has been one of the largest domains of its employment. Often the broad catch-all term *Quantum Information* is used to embrace every aspect of information processing related to Quantum mechanics.

However, in order to prevent us from falling into confusions about terminology we suggest (based on [Nielsen and Chuang, 2000]) to distinguish between *Quantum Information Theory* (QIT) and what may be called *Quantum Information Science* (QIS). Whereas the latter encompasses a wide range of applications such as quantum teleportation and the no-cloning theorem, quantum computation, quantum cryptography, quantum error correction, etc., QIT investigates much more elementary quantum tasks. For the purpose of this thesis, we can then largely es-

chew the applications of QIS and turn our analysis to QIT instead.[1] Despite having refined our object of study, we're still faced with a seemingly obscure subject; as Nielsen and Chuang state

> "quantum information theory may look like a disordered zoo to the beginner, [...] because the subject is under development, and it's not yet clear how all the pieces fit together." [Nielsen and Chuang, 2000, p. 51]

This chapter shall provide a brief introduction to QIT, in order to make the 'disordered zoo' look like an ordered one. Keeping the order of the previous chapters, we start with the formal details of *Quantum Shannon Information* in section (7.2), followed by *Quantum Kolmogorov Complexity* in section (7.3). But first, we will provide some preliminaries of QIT.

7.1 Basic Quantum Information Theory

The difference between the classical descriptions of the world and Quantum mechanics results from the different properties of classical and quantum states. Classical bits, i.e. distinguishable Boolean states 1 and 0, are generally macroscopic systems (for instance, a wire carrying a binary signal). In quantum mechanics the classical Boolean states 1 and 0 can only be represented by reliably distinguishable microscopic quantum states or *qubit*s $|0\rangle$ and $|1\rangle$.[2] These states are called *computational basis* states and correspond to the analogues of 1 and 0 in the classical case. Indeed, in the quantum case, a distinction with zero probability of error can only be achieved with orthogonal quantum states, e.g. horizontal and vertical photon polarizations: $|0\rangle = \leftrightarrow$ and $|1\rangle = \updownarrow$ (see next paragraph for a more detailed example). However, in contrast to classical mechanics, qubits can also be in states of superposition, mathematically represented as a complex linear combination of two orthogonal quantum states $|0\rangle$ and $|1\rangle$

[1] Remember, the applications of information theory to quantum mechanics, i.e. QIS, are largely successful and rather uncontroversial, whereas the claims about the nature of (quantum) information itself are multitude.

[2] 'Qubit' is taken to be the basic unit of quantum information. The term is constructed from 'quantum' and the classical 'bit', used in Information Theory.

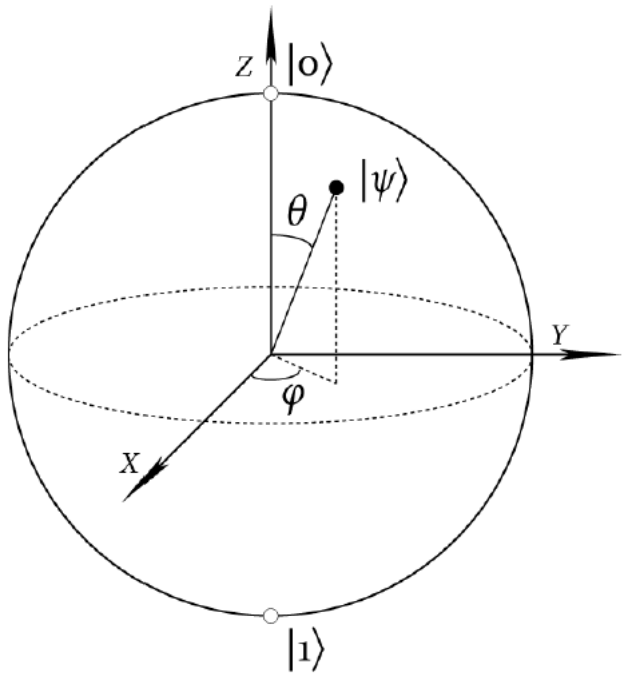

Figure 7.1: Bloch sphere representing the state $|\psi\rangle = \cos\left(\frac{\theta}{2}\right)|0\rangle + e^{i\varphi}\sin\left(\frac{\theta}{2}\right)|1\rangle$, where $0 \leq \theta \leq \pi$ and $0 \leq \varphi \leq 2\pi$.

(for photons e.g., we can find other polarizations such as $\nwarrow\!\!\searrow = \sqrt{\frac{1}{2}}(|0\rangle + |1\rangle)$ or $\circlearrowright = \sqrt{\frac{1}{2}}(|0\rangle + i\,|1\rangle)$). Generalized, a qubit can be written as

$$|\psi\rangle = \alpha\,|0\rangle + \beta\,|1\rangle\,, \tag{7.1}$$

where α and β are complex numbers such that $|\alpha|^2 + |\beta|^2 = 1$. For the different values of the pair α and β there are continuously many states $|\psi\rangle$ a qubit may have. This fact can be visualized in a Bloch sphere Fig. (7.1), where any point on the surface of the sphere represents a pure state. This picture leads to the idea that qubits may represent infinitely more information than classical bits, with their two state space (a bit, represented by a fair coin, may only have head or tails as outcome, represented by $|0\rangle$ and $|1\rangle$ in the Bloch sphere).

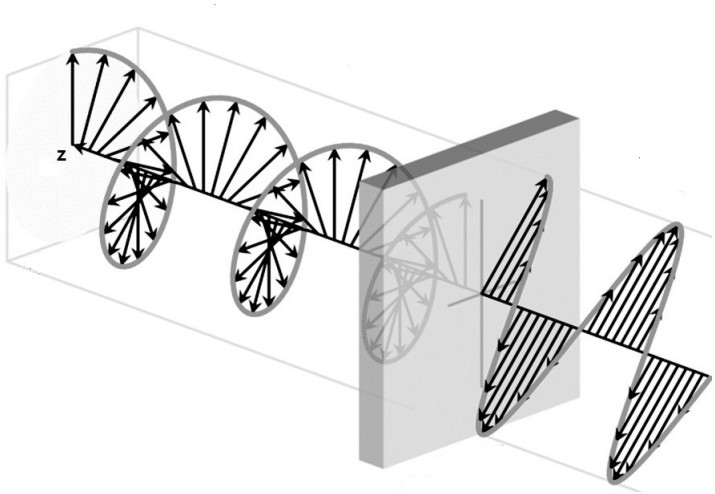

Figure 7.2: Polarized light propagating in z-direction. The arrows represent the electric field $\vec{E}$, oscillating in a plane orthogonal to z.

Polarized light as an example of single qubits.

Let's, for reasons of illustration, regard an example of a practical realization of qubits. In the following, we are using Bub's 'standard example' of polarized light found in [Bub, 2016]. In general we treat light as an electromagnetic wave, generated by the oscillation of electric $\vec{E}$ and magnetic $\vec{B}$ fields in a plane orthogonal in respect to the waves direction of propagation. As seen in Fig. (7.2), light can be polarized according to the direction in which $\vec{E}$ oscillates. Light can then, for instance, be linearly polarized (i.e. when the direction of the oscillating $\vec{E}$ along a fixed direction, e.g. denoted by '$\leftrightarrow$') or circularly polarized (i.e. when $\vec{E}$ rotates during the oscillation, above denoted by '$\circlearrowright$', for instance).

Optical devices, such as polarizing filters, analyzers, and beamsplitters, can be used to transmit only certain components of $\vec{E}$ depending on an angle θ, which denotes the angle of the optical axis of the polarizing filter to the x axis. So for instance, the analyzer transmits only the component of $\vec{E}$ oscillating in the θ direction

$$\vec{E}_\theta = \vec{E}\cos\theta, \tag{7.2}$$

accordingly blocking the orthogonal component

$$\vec{E}_{\theta+\frac{\pi}{2}} = \vec{E}\sin\theta. \tag{7.3}$$

With the insights gained from equations (7.2) and (7.3), and one of the optical devices mentioned above, we are able to prepare single photons in a particular state of polarization.

Moreover, we are now able to practically set up a two-state quantum or qubit system. With a beamsplitter, for instance, an optical device which splits incoming light beams in a 'horizontally' and 'vertically' polarized component, and with an additionally installed photon detector for each beam, we can create a binary-type measurement situation. By applying the *Born rule*, we find that the probabilities for a single polarized photon for leaving the beamsplitter (with the angle θ between the optical axis of the beamsplitter and the polarization of the photon) are

$$p_0 = \cos^2\theta, \tag{7.4}$$

$$p_1 = \sin^2\theta. \tag{7.5}$$

In our case, the subscripts of 0 and 1 shall display that the observables in our set up are associated with binary outcomes.

Entangled states

As we have seen, every quantum state of a system A can be written as

$$|\psi\rangle_A = \sum c_i \, |a_i\rangle \,, \tag{7.6}$$

with respect to an arbitrary set of orthonormal states $|a_i\rangle$ (a basis). In order to describe two or more qubit systems (let's consider two quantum systems A and B), one takes the tensor product of the corresponding Hilbert spaces $\mathcal{H}^{\mathcal{A}} \otimes \mathcal{H}^{\mathcal{B}}$. A general pure state of the compound system AB is then given by

$$|\Psi\rangle_{AB} = \sum c_{ij} \, |a_i\rangle \, |b_j\rangle \,, \tag{7.7}$$

where $|a_i\rangle \in \mathcal{H}^{\mathcal{A}}$ is a basis in $\mathcal{H}^{\mathcal{A}}$ and $|b_j\rangle \in \mathcal{H}^{\mathcal{B}}$ is a basis in $\mathcal{H}^{\mathcal{B}}$. Such a state $|\Psi\rangle_{AB}$is called *separable*, if it can be expressed as a product state $|\psi\rangle_A |\phi\rangle_B$. In case that the coefficients c_{ij} are such that $|\Psi\rangle_{AB}$ can't be expressed as such a product state, then we call the state *entangled*.

7.2 Quantum Shannon Theory

Let's now turn to the equivalent of Shannon's classical information measure in quantum mechanics.[3] Even though quantum information might appear to be fundamentally different than classical information, we can nevertheless find a quite similar framework for the quantum case. The basic idea behind quantum Shannon information is to extend the notion of compressibility to a probabilistically behaving source of qubits. In that respect, Schumacher's [Schumacher, 1995] is a great conceptual breakthrough of QIT, proving the *Quantum noiseless coding theorem*–the equivalent of Shannon's noiseless coding theorem.

Just like in the classical setting, a quantum communication system consists of a quantum signal source S_{QM}, a transmitter, a channel C_{QM} with a quantum signal, a receiver R_{QM}, and a destination D_{QM} (for comparison see section (2.1)). In the first instance, a particular quantum source A,[4] can be thought of as a black box emitting a sequence of quantum systems or signal states (which may be orthogonal or non-orthogonal). Such an emitted string of qubits of length N, then looks like [5]

$$\rho^{\otimes N} = \rho \otimes ... \otimes \rho, \tag{7.8}$$

where ρ describes the density operator

$$\rho = \sum_i p_i \rho_i, \tag{7.9}$$

[3]More (formally) detailed introductions to the topic can be found in [Bub, 2007] and [Nielsen and Chuang, 2000].

[4]From here on we call our particular quantum sources, respectively destinations $'A'$ and $'B'$, in accordance with *Alice* and *Bob*, the usually chosen names in the communication context.

[5]For signal states being prepared in pure states, such a sequence may (more intuitively) look like $|\psi_1\rangle |\psi_1\rangle |\psi_0\rangle ... |\psi_1\rangle |\psi_0\rangle |\psi_0\rangle$.

with $\rho_i = |\psi_i\rangle \langle\psi_i|$ and $\sum_i p_i = 1$.

Similar to the classical case, we now may wonder how much a N-letter quantum message like (7.8) can be compressed. A simple strategy would be to compress a sequence like (7.8) according to the classical $H(X)$. However, taking into account the quantum features of the tokens used in QIT, the optimal compression was found by Schumacher to be

$$\dim \mathcal{H} = 2^{N(S(\rho)+O(1))}, \tag{7.10}$$

where $O(1)$ can be shown to be 'ε', analogous to the considerations of ε-typical messages in context of Shannon's information measure. In the quantum case nearly all sufficiently long messages have support on a *typical subspace* with dimension $2^{NS(\rho)}$, achieving high fidelity by only encoding this typical subspace.[6] In (intuitive) analogy to our example in the previous section, we can then compress a message consisting of N photon states to a number of $W_{typ} = NS(\rho)$ photons. Here $S(\rho)$ denotes the *von Neumann entropy*

$$S(\rho) = -\mathbf{Tr}\rho \log \rho = -\sum_{i=1}^{n} \lambda_i \log \lambda_i, \tag{7.11}$$

where ρ is a density operator on a n-dimensional Hilbert space $\mathcal{H}^n$ and λ_i the corresponding eigenvalues. In fact, in case of pure states, interpreting the eigenvalues λ_i of the density operator ρ as probabilities p_i, leads to $H(X)$ and $S(\rho)$ being formally equivalent.

Let's for a comparison regard some basic properties of $S(\rho)$ [Nielsen and Chuang, 2000]

1. The von Neumann entropy is non-negative and zero *iff* the state is pure.

2. On a n-dimensional Hilbertspace $\mathcal{H}^n$, $S(\rho)$ takes its maximum value $\log n$ for maximally mixed ρ.

3. The Triangle inequality $S(\rho_{AB}) \geq |S(\rho_A) - S(\rho_B)|$ contrasts with the

[6] A second subspace will have a vanishingly small weight of $\rho^{\otimes N}$ as the length $N \to \infty$. This subspace may be seen in analogy to the atypical messages in classical Information Theory.

classical analogue $H(X,Y) \geq H(X), H(Y)$, (compare (2.13)). Whereas in the classical case a bipartite system exceeds the amount of information contained in either part, this doesn't hold in the quantum case. When a composite system AB is in a pure state, then it follows from the first property $S(\rho_{AB})$, that $S(\rho_A) = S(\rho_B)$ (being nonzero if the state is entangled). We can't infer how the state was prepared by observing the two subsystems A and B separately, instead information is encoded in nonlocal quantum correlations.

4. $S(\rho)$ satisfies the inequality $S(\rho) \leq \sum_i p_i S(\rho_i) + H(p_i)$.

Despite some formal similarities between $H(X)$ and $S(\rho)$ (the first two properties), we have to consider some new features (property 3, for instance), stemming from the different nature of quantum sources. In the following we will depict how different alphabets of the quantum source affect the encoding of information.

Pure orthogonal states

First, we may regard $S(\rho)$ as describing an ensemble notion in the quantum realm as an extension of Gibb's classical notion of entropy. In that case, A is thought of as emitting a sequence drawn of the *ensemble* or alphabet $\mathcal{E} = \{|\psi_i\rangle, p_i\}$ of *pure* states distributed according to the probabilities p_i, where $\rho_i = |\psi_i\rangle \langle\psi_i|$. However, *only* in the special case when the signal states $|\psi_i\rangle$ are orthogonal to one another and hence distinguishable, the von Neumann entropy will equal Shannon's measure $H(A) = S(\rho)$. In other words, a qubit encodes exactly one bit and the sequence might as well have been send classically.

Pure non-orthogonal states

The much more interesting 'quantum case' on the other hand, occurs when the signal states are no longer orthogonal and hence not distinguishable. How does this affect the relation between $H(X)$ and $S(\rho)$? The indistinguishability of the non-orthogonal states mirrors the limited

amount of *accessible information* [Schumacher, 1995] when we are dealing with quantum information. In order to determine how much information content we can indeed access, we've to consider the mutual information $H(X:Y)$ (2.18), quantifying how much having received Y helps us to infer about X. In general, the maximum of such accessible information is given by the maximum over all possible measurement schemes (POVM)

$$Acc(\mathcal{E}) = \max H(X:Y). \tag{7.12}$$

A significant theorem by Holevo, the so called *Holevo Bound,* provides an upper bound for $\max H(X:Y)$ on quantum channels

$$H(X:Y) \leq \underbrace{S(\rho) - \sum_i p_i S(\rho_i)}_{\equiv \chi(\mathcal{E})}, \tag{7.13}$$

where $\chi(\mathcal{E})$ is the so called *Holevo Information* (sometimes also referred to as Holevo Chi). With the fourth property of $S(\rho)$, we then obtain

$$H(X:Y) \leq S(\rho) - \sum_i p_i S(\rho_i) \leq H(X). \tag{7.14}$$

Before analyzing the important implications of this expression, let us have a quick a look at another possibility of encoding quantum sequences.

Mixed states

The following quantum communication scheme is sometimes referred to as 'entanglement notion'. In this case, the letters are drawn from the ensemble $\mathcal{E} = \{\rho_i, p_i\}$, such that our quantum source A contains mixed states.

Once again though, we need to differ between orthogonal and non-orthogonal states. If the letters are drawn from an ensemble where the mixed states are mutually orthogonal, such that

$$\mathbf{Tr}\rho_i\rho_j = \delta_{ij}, \tag{7.15}$$

then these states are also perfectly distinguishable. In that case we can essentially conceive the messages as classical and the possible compression reduces to $H(X)$ qubits per letter once again. We could then, e.g., extend our Hilbert space $\mathcal{H}_A$ of our mixed states to the larger space $\mathcal{H}_A \otimes \mathcal{H}_B$, creating the composite system AB (with the 'auxiliary' system B) and choose a *purification*[7] of each ρ_i, namely a pure state $|\psi\rangle_{AB}$, such that

$$\mathbf{Tr}_B \left(|\psi\rangle_{AB} \langle\psi|_{AB} \right) = (\rho_i)_A. \tag{7.16}$$

These pure states $|\psi\rangle_{AB}$ are mutually orthogonal, such that the updated alphabet $\{|\psi\rangle_{AB}, p_i\}$ can be compressed according to the classical case $H(X)$ again.[8] Since ρ_i are non-orthogonal $S(\rho_i) \neq 0$, the third property of the von Neumann entropy, tells us that the states are entangled–that's why one might call it entanglement notion. In case where a message is transmitted with non-orthogonal mixed states, it's impossible to compress such a message with $S(\rho)$, but only with the Holevo Information

$$\chi(\mathcal{E}) = S(\rho) - \sum_i p_i S(\rho_i). \tag{7.17}$$

Holevo-, Accessible- and Specification Information

Reconsidering the inequalities (7.14), we have to take a closer look at the concept of Holevo-, accessible and *specification* information again (the latter term is coined by [Timpson, 2013]). Overall, we can regard the Holevo Information $\chi(\mathcal{E})$ (7.17) as a generalization of the von Neumann entropy and therefore also of Shannon's $H(X)$. For pure states we obtain equality $\chi(\mathcal{E}) = S(\rho)$, since $S(\rho_i) = 0$ for pure states; for orthogonal states we have seen $H(X) = S(\rho)$, so in this case $\chi(\mathcal{E}) = H(X)$. For non-orthogonal mixed states, none of these equalities holds. Because of that, it is fruitful to analyze inequality (7.14) again. With the definition of accessible information (7.12), the Holevo bound (7.13) and the Holevo Information

[7] Purification describes the case that every mixed state ρ_i acting on a finite-dimensional Hilbert space can be regarded as the *reduced* state of some pure state of a larger Hilbert space.

[8] While decoding the state, we can perform the partial trace Tr_B by neglecting subsystem B, and so reconstruct the original message.

(7.17), we obtain

$$Acc(\mathcal{E}) \leq \chi(\mathcal{E}) \leq H(X). \tag{7.18}$$

The unique features of quantum states, i.e. being possibly mixed or non-orthogonal and thus not always being perfectly distinguishable, then forces us to differentiate between (the already introduced) accessible information $Acc(\mathcal{E})$ and specification information. Whereas the former describes the amount of information encoded into sequences of qubits that can also be retrieved from such a message again (in a perfect scenario), the latter enumerates the amount which is necessary for *specifying* qubit sequences.

In a classical context, the amount of specification information never exceeds the amount of accessible information, in fact, they always coincide; a typical bit-strings of $\{1, 0\}$ are (disregarding noise) always completely accessible and the information how to specify them, is given by $NH(X)$. This explains why we didn't have to bother about the distinction between accessible information and specification information when introducing classical information theory in the previous chapters.

However, as we have seen above, the situation is tremendously different in the quantum case. Here equality $Acc(\mathcal{E}) = H(X)$ merely holds *iff* orthogonal quantum states are used for transmission. This can either be achieved for orthogonal states in system A, or for the 'entanglement notion' with a larger system AB, where the subsystem transmits mixed orthogonal states. So just in the special case when orthogonal states are sent can we achieve that (like in the classical case) the amounts of specification information and accessible information coincide. In general, we therefore can't encode more than one single classical bit into each qubit.

In all other cases, the Holevo information will bound the accessible information from above (that's essentially what the Holevo bound (7.13) conveys), such that also

$$Acc(\mathcal{E}) < H(X) \tag{7.19}$$

holds. In other words, in principle it won't be possible to distinguish non-orthogonal states perfectly and hence in a communication scenario

decoding at the destination won't yield an unambiguous message so that $H(X) \neq \chi(\mathcal{E})$. In converse, that means that an arbitrarily large amount of classical information can be encoded in one qubit, without being able to access it though.

7.3 Quantum Kolmogorov Complexity

With some delay in comparison to Shannon's formalism, the notions of Kolmogorov Complexity are slowly transferred into the quantum realm. The main idea of *Quantum Kolmogorov Complexity* is to denote some measure of complexity of an individual quantum state $|\psi\rangle$. Contrary to the classical case, a quantum state can be in superposition, such that e.g.,

$$|\psi\rangle = \frac{1}{\sqrt{2}} |001\rangle + |11010\rangle \,, \tag{7.20}$$

compare the general case (7.6). The complexity $C(|\psi_i\rangle)$ of the quantum state is then defined as the shortest program that describes that state on a universal quantum computer (universal quantum TM) that defines $|\psi\rangle$.

However, doing so is not straightforward and requires some further scrutiny. First we have to fix a basic notion of (universal) Quantum Turing Machines ($\mathcal{QTM}$s). Note that defining a $\mathcal{QTM}$ doesn't coincide with overcoming the technical obstacles of actually constructing a working quantum computer (as stated above, such an undertaking belongs to the category of QIS). Only after the section on $\mathcal{QTM}$s we're ready to devote our attention to the *different* approaches in order to define Quantum Kolmogorov Complexity.

7.3.1 Quantum Turing Machines

As we have seen in Chapter 3, the definition of Kolmogorov Complexity crucially depends on $\mathcal{UTM}$s. One of the main problems of defining Kolmogorov Complexity for the quantum case, is how to implement (universal) $\mathcal{QTM}$s. The main difference to the classical case is that our $\mathcal{QTM}$ will be able to produce and act on linear superpositions of classical con-

figurations.

In 1985 Deutsch [Deutsch, 1985] suggested the first model of a $\mathcal{QTM}$ based on an even earlier proposal by Feynman [Feynman, 1982]. Our presentation in contrast largely relies on [Bernstein and Vazirani, 1997] (and secondary literature there upon, e.g. [Mueller, 2007], [Benatti, 2009]), who refined the theory in more detail.[9] Accordingly, the composition of a $\mathcal{QTM}$ is very much alike to classical $\mathcal{TM}$s, containing:

1. An internal control unit $C_{\mathcal{QTM}}$ with the associated Hilbert space $\mathcal{H}_C$, which is linearly spanned by the orthonormal control states $q \in Q$ describing the state of the control unit.

2. An input/output tape $T_{\mathcal{QTM}}$, with the associated Hilbert space $\mathcal{H}_T$,[10] where σ denotes the content of the tape cells.

3. A read/write head $H_{\mathcal{QTM}}$, with Hilbert space $\mathcal{H}_H$, where k describes the position of the head.

Every $\mathcal{QTM}$ can then be described by means of a Hilbert space

$$\mathcal{H}_{\mathcal{QTM}} = \mathcal{H}_T \otimes \mathcal{H}_C \otimes \mathcal{H}_H, \tag{7.21}$$

with the configuration basis vectors$|\sigma, q, k\rangle$ providing an orthonormal basis.[11] Just like its classical pendant (3.2), a $\mathcal{QTM}$ can be defined as a 7-tuple[12]

$$M_{QM} = \{Q, \Sigma, \Gamma, \delta, q_0, b, \mathcal{F}\}, \tag{7.22}$$

[9]For the more interested reader, a more detailed treatment of $\mathcal{QTM}$ can then be found in the just named sources.

[10]For reasons of convenience, we consider a special class of $\mathcal{QTM}$s, with their tape $T_{\mathcal{QTM}}$ consisting of two different tracks, an *input track* I and an *output track* O, such that $\mathcal{H}_T = \mathcal{H}_I \otimes \mathcal{H}_O$.

[11]The states of the respective Hilbert spaces are $|\Psi_C\rangle = \sum_{i=1}^{|Q|} c_i |q_i\rangle$,with $\sum |c_i|^2 = 1$; $|\Psi_T\rangle = \sum_{\sigma \in \sum^z} t_\sigma |\sigma\rangle$; and $|\Psi_H\rangle = \sum_{k \in \mathbb{Z}} h_k |k\rangle$,with $\sum |k_i|^2 = 1$.

[12]Note that in [Bernstein&Vaziriani Def.3.2] a $\mathcal{QTM}$ is originally defined as $M_{QM} = \{Q, \Sigma, \delta\}$. However, in order to remain consistent with our defintion of classical $\mathcal{TM}$s given earlier, we added $\Gamma, q_0, b,$ and $\mathcal{F}$ to define M_{QM}. Conceptually, the original definition of M_{QM} isn't altered though, we just used a different notation here. $\Gamma, q_0, b,$ and $\mathcal{F}$ can simply be thought of as being included in the triplet$\{Q, \Sigma, \delta\}$.

where the variables are analogous to the classical case, in the sense that the set of states Q and the set of type symbols Γ are each replaced by a Hilbert space, such that b corresponds to a zero-vector and the initial states q_0 are either pure or mixed states. In addition, we can define the *Quantum transition function*

$$\delta = Q \times \Sigma \to \tilde{\mathbb{C}}^{Q \times \Sigma \times \{L,R\}}, \tag{7.23}$$

where $\sum$ expresses the input symbols and $\tilde{\mathbb{C}}$ expresses the set of complex numbers which are efficiently computable.

Comparing the set up of a $\mathcal{QTM}$ to a classical $\mathcal{TM}$, we note that the quantum transition function resembles the probabilistic transition function in the classical case.[13] Instead of having a single classical successor state, a probabilistic transition function chooses between a set of available successor states according to a probability distribution. In a similar fashion, the quantum δ assigns *amplitudes* instead of classical probabilities.

7.3.2 Approaches to define Quantum Complexity

Contrary to the quite straightforward definition of Kolmogorov Complexity in the classical setting, the quantum case (so far) doesn't allow for a smooth and straightforward application of algorithmic information.[14] The main question is how to account for the quantum setting; should we consider a classical or quantum reference machine? Should the input/output tape of such be classical or quantum? Faced with a sheer infinite amount of indistinguishable qubit strings, do we allow for small errors? Answering these questions in different ways and applying simple combinatorics explains why we are faced with many different ap-

[13] We can find such a 'probabilistic transition function' in so called *Probabilistic Turing Machines* $\mathcal{PTM}$, which is a non-deterministic $\mathcal{TM}$ that chooses the successor state according to a probability distribution.

[14] One may argue though, that the absence of a clear formalism for Kolmogorov Complexity in the Quantum case is due to the recentness of the approach. Perhaps the future holds a simple and straight forward solution.

proaches.[15] For an overview of these different approaches see Appendix (10.4). As of now, it is still an open problem how all these approaches relate to one another. A few words about how this affects the goal of the current thesis will follow in the next chapter.

However, in the following we adapt the version of Quantum complexity brought forward by [Berthiaume et al., 2001], as presented in [Benatti et al., 2006], [Mueller, 2007], [Mueller and Rogers, 2008] and [Benatti, 2009]. Among the reasons for choosing this options is that we want our results to be as general as possible and allow for classical and quantum in- and outputs for our $\mathcal{QTM}$. Such a $\mathcal{QTM}$ will naturally produce superpositions of qubit strings of different length, called *indeterminate* or *variable length qubit strings* (like e.g, seen in expression (7.20)) [Mueller, 2007]. Let $\mathcal{H}_k(\mathbb{C}^{\{0,1\}})^{\otimes k}$ be the Hilbert space of k qubits ($k \in \mathbb{N}_0$), with$\mathbb{C}^{\{0,1\}}$written for $\mathbb{C}^2$, as indication that we fix $|0\rangle$ and $|1\rangle$as the computational basis vectors. The Hilbert space containing variable length qubit strings of length k is then denoted by

$$\mathcal{H}_{\{0,1\}} \oplus_{k=0}^{\infty} \mathcal{H}_k, \tag{7.24}$$

with $\mathcal{H}_{\leq n} \oplus_{k=0}^{n} \mathcal{H}_k$ being a subspace of $\mathcal{H}_{\{0,1\}}$.

Defining length

We can then define the length $l(\sigma)$ of a qubit string $\sigma \in \mathcal{T}_1^+(\mathcal{H}_{\{0,1\}})$ as

$$l(\sigma) \min \left\{ n \in \mathbb{N}_0 \mid \sigma \in \mathcal{T}_1^+(\mathcal{H}_{\leq n}) \right\}, \tag{7.25}$$

with $\mathcal{T}_1^+(\mathcal{H})$ denoting the density operators (i.e. positive trace-class operators with trace 1). Moreover, we can define the 'average length' $\bar{l}(\sigma)$ as

$$\bar{l}(\sigma)\mathbf{Tr}(\sigma\Lambda), \tag{7.26}$$

where Λ denotes the unbounded self-adjoint length operator $\Lambda |x\rangle\, l(x)\, |x\rangle$ for all classical strings $x \in \{0,1\}$. The idea to define the average length Quantum Complexities is due to [Rogers and Vedral, 2008], who argue

[15] The following summary is partly based on [Mueller, 2007]; the following definitions of the quantum algorithmic information measures then might slightly diverge from their original formulation (not conceptually though).

that it's correlated to the average energy of the input. In addition, Mueller [Mueller, 2007] argues that the thereof derived average-based notion of Quantum Complexity has the advantage of accounting for applications in statistical mechanics.

However, as we have seen earlier, there exist an infinite amount of qubit strings of indefinite length that aren't perfectly distinguishable, such that determining the exact length is unfeasible. Like in the previous attempts to define Quantum Complexity, this infinite amount of strings motivates us to allow for small errors. Basically, this can be done in two ways. First, we can allow for a certain tolerance $\delta > 0$, using either fidelity or trace distance.[16] Second, we can provide our $\mathcal{QTM}$ with a parameter $k \in \mathbb{N}$ and demand an accuracy of the output to k digits. As pointed out in [Mueller, 2007], the second procedure is analogues to a classical algorithm that calculates $\pi = 3.14...$ to the kth digit.

Quantum Complexities

With two different notions of length (7.25) and (7.26), and two options to account for errors we obtain no less than four possible definitions of Quantum Complexity

$$QC^{\delta}_{\mathcal{U}} \min \left\{ l(\sigma) \ : \ \| \rho - \mathcal{U}(\sigma) \|_{\mathrm{Tr}} < \delta \right\}, \tag{7.27}$$

$$\overline{QC}^{\delta}_{\mathcal{U}} \min \left\{ \bar{l}(\sigma) \ : \ \| \rho - \mathcal{U}(\sigma, k) \|_{\mathrm{Tr}} < \delta \right\}, \tag{7.28}$$

$$QC_{\mathcal{U}} \min \left\{ l(\sigma) \ : \ \| \rho - \mathcal{U}(\sigma, k) \|_{\mathrm{Tr}} < \frac{1}{k} \ for \ every \ k \in \mathbb{N} \right\}, \tag{7.29}$$

$$\overline{QC}_{\mathcal{U}} \min \left\{ \bar{l}(\sigma) \ : \ \| \rho - \mathcal{U}(\sigma, k) \|_{\mathrm{Tr}} < \frac{1}{k} \ for \ every \ k \in \mathbb{N} \right\}. \tag{7.30}$$

[16] Instead of the fidelity, as originally used in [Berthiaume et al., 2001], the trace distance is used in the following presentation.

Let us add a few remarks. As described in [Benatti, 2009, §9], the exact specification $\frac{1}{k}$ isn't important as long as $f(k)$ is a computable function such that $f(k) \to 0$ for $k \to \infty$. Furthermore, the choice of the code C for encoding of the variables σ and k is irrelevant (up to some constant), as long as the variables can both be computably decoded from $C(\sigma, k)$.

Quantum Invariance Theorem

Analogously to the classical case, we can formulate the Invariance Theorem (3.6) for the quantum case (given $\mathcal{U}$ is our reference machine)

$$QC_{\mathcal{U}}(\rho) \leq QC_M(\rho) + c_M, \tag{7.31}$$

for every quibit string ρ.[17] Choosing a different reference machine $\mathcal{U}$ will change the value of $QC(\rho)$ only up to a constant.

Quantum Kraft Inequality

Additionally, Mueller and Rogers [Mueller and Rogers, 2008] showed that there exists a quantum version of the Kraft inequality (4.6), holding for arbitrary prefix-free Hilbert spaces (even if they don't posses an orthonormal basis of length eigenstates). Such a quantum version then reads as

$$\sum_{i \in I} 2^{-l(e_i)} \leq \sum_{i \in I} 2^{-\bar{l}(e_i)} \leq \mathbf{Tr}\left(2^{-\Lambda}\mathbb{P}(\mathcal{H})\right) \leq 1, \tag{7.32}$$

where $\{|e_i\rangle\}_{i \in I} \subset \mathcal{H}_{\{0,1\}}$ is a prefix-free orthonormal system, spanning in the Hilbert space $\mathcal{H} \subset \mathcal{H}_{\{0,1\}}$, with equality for the three left terms if and only if every $|e_i\rangle$ is a length eigenstate.

Quantum Complexity of Classical Strings

At last, it can be demonstrated that quantum Kolmogorov Complexity 'extends' classical complexity in a similar way as von Neumann entropy generalizes Shannon entropy [Mueller, 2007], such that

[17] Essentially the same relation can be shown for $QC^{\delta}_{\mathcal{U}}$ (7.27). Whether the Invariance Theorem holds for the averagle length cases (7.29) and (7.30) is

$$QC(|x\rangle) + c = K(x), \tag{7.33}$$

with some constant $c \in \mathbb{N}$ and where $K(x)$ denotes the classical Kolmogorov Complexity (3.5) of a classical sequence $x \in \{0,1\}$. We then obtain relation (7.33) by choosing the (classical) reference machine of $K(x)$ to be a reversible one. Since every reversible $\mathcal{TM}$ is also a (special case of a) $\mathcal{QTM}$, we can then apply the *Quantum Invariance Theorem* (7.31) such that the above relation follows [Mueller, 2007].

7.4 Quantum Brudno theorem

Much like we've already seen in chapter 4, we can establish a connection between Quantum Shannon Information and Quantum Algorithmic Information. In contrast to the classical case, where the different information measures are 'directly' related, the *Quantum Brudno theorem* connects the quantum entropy *rate* s (of an ergodic quantum spin chain) to the qubit complexity $\frac{1}{n}QC(\rho)$.[18] The notion of entropy- or complexity rates is often useful to compare sequences of different lengths. In general, the entropy rate of a (stochastic process) is defined as the limit of the joint information measure (2.13)

$$h(X_i) = \lim_{n\to\infty} \frac{H(X_1, X_2, ..., X_n)}{n}. \tag{7.34}$$

Note that in case of a Bernoulli type information source, the entropy rate will coincide with $H(X)$. Next to $h(X)$, we can also define a complexity rate as

$$k(x) = \lim_{n\to\infty} \frac{1}{n} K(x^{(n)}), \tag{7.35}$$

where $K(x)$ is our well-known Kolmogorov Complexity (3.5) and $x^{(n)}$ denotes the first n digits of an infinite binary sequence x. Conjectured by Zvonkin and Levin [Zvonkin and Levin, 1970] and proved by Brudno

[18] The reader may cautiosly interpret the term 'entropy' appearing in this subsection as our notion of (syntactic) information.

[Brudno, 1982], it can be shown that

$$k(x) = h(X)$$

for almost all sequences, where $h(X)$ is supposed to denote the entropy rate of a binary ergodic source and $k(x)$ the algorithmic complexity rate of trajectories in the phase space. Recently it was shown in [Benatti et al., 2006], that the classical result can be adapted to the quantum case. Defining the entropy rate of a quantum information source as

$$s(\Psi) = \lim_{n\to\infty} \frac{1}{n} S(\rho^{(n)}), \tag{7.36}$$

where S denotes the von Neumann entropy (7.11) and Ψ a state of an ergodic quantum source $(\mathcal{A}^{\infty}, \Psi)$. The *Quantum Brudno Theorem* then states that the quantum complexity rates are constrained by an upper and lower bound,[19] such that

$$\frac{1}{n} QC^{\delta}(q) \in (s - \delta(4+\delta)s, s + \delta), \tag{7.37}$$

$$\frac{1}{n} QC(q) \in (s - \delta, s + \delta), \tag{7.38}$$

where s denotes the entropy rate of the quantum source and δ a certain tolerance. The Quantum Brudno Theorem hence provides the link to relate our quantum notions of Shannon Information and Kolmogorov Complexity.

7.5 Summary

At the beginning of the chapter we suggested to distinguish between Quantum Information Science (QIS) and Quantum Information Theory (QIT). While the former gives rise to many new potential applications, the latter is crucial for our analysis of different notions of information in respect to their ontology.

[19] The lower bounds are given by $\frac{1}{n} QC^{\delta}(q) > s - \delta(4+\delta)s$ and $\frac{1}{n} QC(q) > s - \delta$; the upper bounds are given by $\frac{1}{n} QC^{\delta}(q) < s + \delta$ and $\frac{1}{n} QC(q)$

In the case of Quantum Shannon Information, we have seen that the same *concepts* of the classical case have been applied to the quantum case. Instead of a classical communication system, we're faced with a quantum communication system, i.e. we replaced a classical source, emitting classical states with a quantum source, emitting quantum states. Due to the fact that quantum states have features that classical states can't have (e.g., being mixed or in superposition), we introduced the so called Holevo Information $\chi(\mathcal{E})$ (7.17) as new compression rate. Nevertheless, we can at most encode one classical bit for one qubit.

However, faced with a large number of different approaches in the quantum case, Quantum Complexities are still very much in the spirit of classical case, such that the main *concepts* haven't changed. Basically, the goal is to find the shortest program that generates a qubit string on a quantum computer. We introduced the notion of a universal Quantum Turing Machine, 'translated' the classical concept of length to the quantum case and derived a number of Quantum Complexities (7.27)-(7.30). Subsequently we could show that many of the characteristics of classical Kolmogorov Complexity, like the Invariance Theorem or the Kraft Inequality, have in fact quantum analogues.

At last, we demonstrated that the notions of quantum Kolmogorov Complexity rates coincide with the von Neumann entropy rate of ergodic quantum information sources.

Chapter 8

Interpreting Quantum Information

> "Quantum information is a new concept with no classical analog, and it is important to distinguish it from the state identity. [...] In more formal terms, we would aim to formulate and interpret quantum physics in a way that has a concept of information as a primary fundamental ingredient."[Jozsa, 2004, p. 79]
>
> – *Josza*

In the previous chapters, we have laid the foundation to finally answer the question of *What's the ontological status of information in physics*? First, introducing the mathematical formalism of Shannon information and Kolmogorov Complexity helped us interpreting each of the respective information measures in the classical case and lay the foundation of this chapter. As we shall argue in the following, much of our analysis from the classical case can be applied to quantum information as well.[1] To do so, we briefly characterize Quantum Shannon Information and Quantum Kolmogorov Complexity and point out the most important differences to the classical case. Analogously to the previous chapters, we then continue by dismissing any relation with uncertainty and semantic information, and apply the type/token distinction. Thereafter, we ana-

[1] Since the lion share of work has been done in chapter 5 and 6 already, both quantum Shannon information and quantum Kolmogorov Complexity are treated in the same chapter.

lyze to what extent the notions of quantum information are conventional. Finally, we draw a conclusion to what degree both quantum information measures are compatible.

8.1 A few words about Quantum Shannon Information

The difference between classical Shannon Information and Quantum Shannon Information emerges due to the different properties of classical and quantum states; only orthogonal quantum states are reliably distinguishable with zero probability, for instance. The aim of Quantum Information Theory is then to determine the optimal rate at which quantum sequences can be communicated, considering the statistical characteristics of a communication system with quantum properties. As we have seen in section (7.2), for encoding a qubit sequence of length N, we require a Hilbert space with minimum dimension 2^N. According to the Schumacher coding theorem [Schumacher, 1995], very much alike to the classical case, Quantum Shannon Information characterizes the best rate of *compression* solely based on the given quantum and statistical properties of the system. If ρ denotes the average density operator of a quantum source, we then only need $NS(\rho)$ qubits (in case the sequence consists only of pure states) to specify our quantum message with high fidelity in a $2^{NS(\rho)}$-dimensional space.

The Relation of Classical and Quantum

While the concept behind Quantum Shannon Information is based on the insights of classical Information Theory, various scholars have come to different conclusions about the relation between the classical and quantum concepts. Some scholars embrace the view (e.g., compare Richard Josza's statement in the epigraph) that because quantum information displays many new features that are unlike to the classical case, an entirely new concept is needed. On the other hand, one might conclude that

'quantum information doesn't exist' (compare to [Duwell, 2003])[2] and argue that Quantum Information is merely a 'synonym for an old concept', in the sense that Quantum Information is basically Shannon's classical theory operating with quantum states.

In this thesis we *sympathize* with the reverse position, namely that classical information should instead be regarded as a subcategory of quantum information;[3] Shannon Information is a special case of Quantum Information, operating with (classical) orthogonal states. Our view is corroborated by our findings of the previous chapter, where we saw that the classical Shannon Information $H(X)$ is equal to the Holevo Information $\chi(\mathcal{E})$ (7.17) iff orthogonal states are used to encode Shannon Type Information. In all the other cases where the states aren't orthogonal, $\chi(\mathcal{E})$ has to be chosen as the generalized information measure.[4]

Although the notion of Quantum Information in a communication setting is ultimately based on Holevo Information–and one might claim to have, as Josza claimed, a new concept after all–our extensive analysis of the classical case shall tremendously help us to understand the nature of Quantum Shannon Information. As it turns out, we can essentially repeat our arguments about Quantity and Type Information in the following sections.

Quantity and Type Information

The notion of Quantity Information in the quantum case is dealt with quickly; we then may use the exact same description of Shannon Quantity Information of section (5.4). In other words, it's irrelevant for Quantity Information whether the tokens of the regarded communication sys-

[2]As pointed out in [Duwell, 2008], the author has since then stepped back from his view.

[3]Of course this view is not entirely new. Jeffrey Bub, for instance, also holds such a position [Bub, 2007], [Bub, 2012]. As well see in section (8.5.1) though, the difference in dependence on conventional elements in the classical and quantum case might raise some questions about the view that Quantum Information is a generalization of Shannon Information

[4]Of course one can also use the von Neumann entropy $S(\rho)$, as long as one doesn't use mixed non-orthogonal states. As seen in $\chi(\mathcal{E}) = S(\rho) - \sum_i p_i S(\rho_i)$ (7.17), the von Neumann entropy is a special case of $\chi(\mathcal{E})$ and obtained when $\sum_i p_i S(\rho_i) = 0$.

tem are classical or bear any extra quantum properties. After all, the Shannon Quantity Information remains an abstract entity in the quantum case too.

But what about Type Information, don't the properties of 'quantum tokens' render our original analysis useless? Not at all! Even though we could only partially introduce the admittedly intriguing features of Quantum Shannon Information ('entanglement', 'teleportation', etc., to only drop a few key words), these features don't affect our previous argument drawn in the classical case. The Schumacher coding theorem assures that Quantum Information is still an optimal compression scheme for coding. As such, we're now entitled to use 'quantum tokens' to encode our messages. So instead of classical tokens (or orthogonal states) we use qubits to generate a sequence which encodes our Quantum Type Information. One may then choose from a variety of 'quantum tokens' to instantiate a certain sequence type–e.g., polarized photons (as seen in the example of previous chapter), the spin of electrons, etc.. Despite often not being able to identify which sequence was sent (see section (7.2)), the basic argument from the notions of Coding Theory doesn't change; the multi realizability of types still allows us to constitute Type Information in multiple ways by the means of tokens. We thus conclude that Quantum Type Information is an abstract entity too.

8.2 A few words about Quantum Complexity

Let's start this section with a disclaimer. Among the fairly recently established field of Quantum Information Theory, Quantum Complexity is among the even more recent developments. The rest of the analysis should therefore be taken with the grain of salt that the notion of Quantum Complexity is more prone to changes than any other notions of well established theories.

The Relation of Classical and Quantum

Like in the case of Shannon Information, the properties of qubit strings forced us to adapt Kolmogorov Complexity to the quantum case. Most notably we had to introduce the notion of universal $\mathcal{QTM}$ and explain how to account for indeterminate qubit strings. So instead of determining the shortest program p to generate a classical bit string x consisting of $\{0, 1\}$, the target of Quantum Complexity are uncountably many qubit strings like $|x\rangle = \frac{1}{\sqrt{2}} |001\rangle + |11010\rangle$. Since we lack the resources to calculate each qubit string $|x\rangle$ with infinite precision, we had to grant the description of the state a certain margin of error.

However, we don't have to conceive Quantum Complexity as an entirely new information measure. Aside from using the same basic idea, we've seen that $QC(|x\rangle) + c = K(x)$ (7.33) holds, such that the notions of Quantum Kolmogorov Complexity coincide with classical $K(x)$ for classical strings. Since this view suggests that $QC(|x\rangle)$ can 'handle' both classical and qubit strings (whereas $K(x)$ can merely handle the former), we therefore *might* regard Quantum Complexity as a generalization of classical Kolmogorov Complexity.

Quantity and Type Information

Just like in the case of Quantum Shannon Information, we can essentially use the same description of Quantity Information for Quantum Complexity (save for replacing 'sequence x' with 'state $|x\rangle$'). Switching from classical to quantum strings doesn't change the way we regard Quantity Information. Algorithmic Quantity Information remains an abstract entity in the quantum case.

In the same vein, our original argument about the abstractness of Algorithmic Type Information doesn't suddenly change because we consider qubit strings $|x\rangle$ instead of classical strings x. While the tokens such strings are composed of certainly have changed (and especially their respective properties), the main argument about the encoding of Algorithmic Type Information in such sequences hasn't. In the quantum case, Algorithmic Type Information is instantiated by multiple realizable qubit

strings; Quantum Algorithmic Type Information *remains an abstractum*.

8.3 Quantum Information As a Measure of Uncertainty?

In this section we argue that neither Quantum Shannon Information nor Quantum Complexity are suitable as measures of uncertainty. In both cases, the arguments basically work as already seen in Chapter 5 and Chapter 6.

Quantum Shannon Information

The 'easiest' argument against Quantum Shannon Information as a measure of uncertainty is to simply copy our argumentation of the classical case: There exists a whole class of measures of uncertainty $U_r(P, \mu)$ (5.1) and only for $r = 0$ (and taking the logarithm) we might obtain $H(X)$. However, failing to find an additional axiom to pick out $H(X)$ as a unique uncertainty measure, leads to a non-tolerable discrepancy with $H(X)$ as a uniquely derived information measure.

Note that at first, it is questionable whether or not our new quantum formalism falls into $U_r(P, \mu)$ for any value of r, in order to qualify as a measure of uncertainty. But even if our new formalism would fall into the class of $U_r(P, \mu)$, it actually doesn't matter since that doesn't provide us with a plausible uniqueness requirement. The classical argument thus still holds.

Quantum Complexity

Quantum Complexity doesn't qualify as a measure of uncertainty either. Quantum Complexity isn't based on what's key to Uffink's concept of $U_r(P, \mu)$, i.e. being based on the notion of probability distributions. In other words, we won't be able to find a value for r to pick out Quantum Complexity.

8.4 Semantics in Quantum Information?

In Chapter 5 and Chapter 6 we argued that in the classical case, neither Shannon Information nor Kolmogorov Complexity are concerned with any aspects of semantic information. Essentially, the very same arguments of either information measure in the classical case, can be repeated in the quantum case.

No Semantics in Quantum Shannon Information

Our original argument against semantic information in the Shannon formalism was based on two observations. First, virtually any type of tokens can be used for transmitting a message in a communication system. The somewhat exotic example of an alphabet consisting of fruit and vegetables was supposed to illustrate that the tokens themselves are meaningless. Second, we argued that what's encoded in a sequence might be meaningful, but that $H(X)$ has no means to denote any value for measuring semantic information content.

Turning to the quantum case, we argued that the Schumacher coding theorem is very much in the same vein as Shannon's coding theorem, describing the minimal resources needed to transmit quantum information. Instead of classical tokens, quantum information is based on a sequence of quantum states (or string of qubits). The source S_{QM} either emits single pure states or quantum states which are parts of larger entangled systems. However, by exchanging classical states with quantum states, the original arguments from the classical case aren't altered at all! Whether the tokens for transmitting messages are concrete letters, electronic impulses, fruits and vegetables or quantum states, doesn't matter. Admittedly, quantum states have some features which won't come along with classical objects (superposition, for instance), but none of these features plausibly account for semantics.

Similarly, replacing 'classical' tokens with quantum tokens, doesn't establish a connection to semantic information. The formalism of Quantum Shannon Information isn't suitable to measuring meaning either.

No Semantics in Quantum Kolmogorov Complexity

Like in the case of Quantum Shannon Information, the 'transition' of Kolmogorov Complexity to the quantum case doesn't change our original arguments against semantic elements. Our notions of quantum complexities don't allow to quantify semantic information either. Analogously to ordinary Kolmogorov Complexity, the syntactic features of a string may allow us to find a short algorithm operating on a $\mathcal{QTM}$ to generate that string. The major differences to the classical case are the inclusion of quantum sequences and $\mathcal{QTM}$s as reference machine. But the quantum characteristics of strings or reference machines are inept to account for semantic information. Quantum complexity hence isn't related to semantic information.

8.5 Is Quantum Information conventional?

So far, basically all of our conclusions from Chapter 5 and 6 also hold for the notions of Quantum Information. As we'll argue in the following, the main differences (relevant for the ontological status of information) between classical and Quantum Information are manifested by a different degree of conventionality. Remember, previously we argued that Shannon Information and Kolmogorov Complexity are almost completely stipulated by those who set up a communication system or determine how to extract the patterns. To quite an extent, Quantum Theory constrains us in our completely conventional choices.

8.5.1 Quantum Shannon information

To avoid the completely conventional elements of the classical case, we have to find criteria that fix the experimental set up and the experiment, and provide a 'natural' success criterion.

Quantum Fidelity

Whereas the success criteria of classical Shannon Theory (see section (5.5)) offer us no clue on how to constrain what counts as successful communication, the situation is different for quantum communication. The relation between the theoretical insights of Quantum Theory and the corresponding experimental procedures allow us to exploit the quantum properties of the states emitted by the source, to find 'natural constraints'. These constraints are 'natural' in the sense that they're dictated by the predictions of Quantum Theory. Rejecting these constraints would be rejecting the empirical success of Quantum Theory.

Duwell [Duwell, 2008] advocates that the so called *entanglement fidelity*–a special case of 'regular' quantum fidelity–is suitable to show that for quantum communication to be successful, the (quantum) states reproduced at the source have to some extent behave like the states emitted from the quantum information source. Based on Uhlmann's transition probability formula [Uhlmann, 1976], Josza [Jozsa, 1994] proposed a fidelity for mixed quantum states to be

$$F(\rho_1, \rho_2) = \left(\mathbf{Tr}(\sqrt{\rho_1}\rho_2\sqrt{\rho_1})^{\frac{1}{2}}\right)^2, \tag{8.1}$$

where ρ_1 and ρ_2 denote the density matrices of two quantum states respectively. A helpful interpretation of the above expression can be found in [Nielsen and Chuang, 2000, §9.2]: The square-root quantum fidelity, equals the square-root fidelity[5]

$$\sqrt{F(\rho_1, \rho_2)} = \min_{\{E_m\}} \sqrt{F(p_m, q_m)} \tag{8.2}$$

of the probability distributions induced by the best discriminating measurement between the states ρ_1 and ρ_2. The probability distributions for

[5]In [Miszczak et al., 2008] it is pointed out that the definition/notation of quantum fidelity F is slightly ambiguous (remember the quote in the introduction of Chapter 7, in which Quantum Information Theory is compared to a 'disordered zoo'). The presentation of quantum fidelity in [Duwell, 2008] for instance, basically relies on [Nielsen and Chuang, 2000, §9.2]. However, the definition of Nielsen and Chuang slightly diverges from the original ones given in [Uhlmann, 1976] and [Jozsa, 1994]. In [Nielsen and Chuang, 2000] the quantum fidelity is defined as $\sqrt{F}$ instead of eq. (8.1), which is based on Uhlmann's and Josza's publication.

these states are $p_m = \text{Tr}(\rho_1 E_m)$ and $q_m = \text{Tr}(\rho_2 E_m)$ corresponding to a POVM measurement $\{E_m\}$. The fidelity is bounded between 0 and 1, obtaining the maximum value when the distributions are identical. Based on *Uhlmann's theorem*, the fidelity is the maximum overlap of all possible purifications of the two states ρ_1 and ρ_2.[6] According to expression (8.2), squaring the overlap provides us with the probability that the two compared purified systems pass a test whether they are the same or not [Duwell, 2008].

Since the purification involves the entanglement of our original system A with an auxiliary system B, it is useful to define the notion of entanglement fidelity $F(\rho, \Delta)$, where quantum system AB is prepared in state ρ and Δ denotes a quantum operation, a process that may describe the transmission in a quantum communication scenario. Based on eq. (8.1), the entanglement fidelity provides us with a measure of how well the entanglement is preserved and can then be used as a success criterion. By choosing a particular δ, we can fix the bound of acceptable fidelity $\geq 1 - \delta$. Based on the above insights it's guaranteed that a length N pure state ρ (a sequence such as $|\psi_1\rangle\,|\psi_1\rangle\,|\psi_0\rangle\,...\,|\psi_1\rangle\,|\psi_0\rangle\,|\psi_0\rangle$) will have a probability $\geq 1 - \delta$ to be obtained in a measurement in the computational basis at the destination.

Overall, ensuring that the entanglement fidelity is high, not only yields a preservation of local behavior of the systems that compose a sequence, but also preserves any entanglement present. Unlike in the classical case, quantum theory hence offers an entanglement based constraint for a success criterion. We conclude that choosing a success criterion is *no longer largely conventional*.

The experimental set up and the experiment

The argumentation above already partially illuminated that QIT sets boundaries on possible quantum communication systems and the tokens used in a communication scenario. In contrast to classical Shannon In-

[6] When the states are pure, an overlap is considered to be $\sqrt{F(\rho_1, \rho_2)} = \left(\sqrt{\langle\rho_1 \mid \rho_2\rangle \langle\rho_2 \mid \rho_1\rangle}\right)^{\frac{1}{2}} = (|\langle\rho_1 \mid \rho_2\rangle|)^{\frac{1}{2}}$. For 'purifications' see (7.16).

formation, we can no longer use any kind of tokens to encode information, we have to rely on 'quantum tokens'. Regarding the *experiment*, we can of course still conventionally decide how to partition the outcomes of the source. However, we can no longer stipulate any kind of elementary events by the choice of an *experimental set up*. A quantum information source ought to emit the quantum tokens that can (at the most basic level) be described as density operators on a Hilbert space and 'behave' according to quantum theory; the choice of the experiment is fixed to the extent that the finest possible outcomes have to be describable by quantum theory. What elementary events can be identified within the states produced by a quantum information source is hence *no longer completely dependent on the user.*

However, as Duwell recently pointed out [Duwell, 2017], the constraints set by Quantum Theory might have a bearing on our view that Quantum information is in fact a generalization of classical Shannon Information (as advocated earlier). Since the latter is completely free of any constraints (whereas Quantum Information is not), we might encounter the situation where the transfer of Shannon Information doesn't entail the transfer of the quantum analogue. It appears therefore puzzling whether the quantum version can indeed be regarded as a generalization of the classical one.

8.5.2 Quantum Kolmogorov Complexity

Let's now investigate to what extent Quantum Complexity suffers from absence of an effective description method. In the classical case, $K(x)$ only displays an intersubjective measure of algorithmic information as far as the sequence x under scrutiny is somehow a priori given. But in order to determine the complexity of objects an effective description method is needed. On top of this, such a description method requires a certain coarse graining to be unambiguous; overall a seemingly hopeless task in the classical case.

In the case of Quantum Kolmogorov Complexity however, we seem to be much more restricted in the first place. The introduction of $QC(|x\rangle)$

was intended as a measure of the descriptions of quantum states only and thus doesn't have to account for descriptions of classical macroscopic 'objects' like apples, paintings or the weather.

Description Method & Coarse Graining

Much more important, quantum theory provides us with a successful rigid formalism that allows us to describe quantum states. Basically, a light version of this formalism has been presented in the previous chapter, where we saw the general notion of a quantum states (see eq. (7.6)) and provided polarized light as an example of single qubits. So unlike the classical case, where we are completely clueless of how to extract patterns from the myriad of possible macroscopic objects, the insights of quantum theory force us to refer to quantum states in a certain predefined way. We can then no longer *fine grain* at will, like in the continuous classical case. At the most fundamental level we're limited by the precision of our measurement apparatus and Heisenberg's uncertainty principle–we can't choose 'finer' elementary events. However, unlike the classical case, the stipulation of a computational basis seems to affect Quantum Algorithmic Information to some degree. While the Quantum Brudno Theorem (see section (7.4)) ensures that QC is upper and lower bounded, thus constraining Quantity Information, the effects on Type Information are yet unsolved. Note in addition, that we're in principle still able to stipulate whatever kind of coarse graining we deem reasonable, save for that at a larger scale the described object might no longer be practically conceivable as a quantum state.

So, in other words, by adhering to the already given framework of quantum theory, we have to–to some extent–rely less on some of the conventional context-dependent choices of the classical case. By restricting our description to 'quantum tokens' and setting a lower bound for our 'resolution', quantum theory dismantles a good portion of the conventional elements of classical Kolmogorov Complexity. Of course quantum theory in itself is not free from featuring some arbitrary conventions–but in the end no scientific theory is! To some extent virtually every scientific theory has to rely on conventions (e.g., how do we set up a coordinate system,

etc.). But these kinds of conventions were not part of our worries about conventionality anyway. Initially we were concerned about how to establish a successful description method that may extract patterns from all kinds of (macroscopic) objects, moreover specifying which degree of coarse graining ought to be appropriate. On the other hand, Quantum Theory is a well established theory, corroborated by enormous empirical success and as such it offers us a stable framework (one that is absent in the classical case) to rely on. Rejecting the description method of quantum theory would therefore amount to rejecting one of our most successful scientific theories over the reasons of some conventions.

8.6 Conclusion

In the end of chapter 6, we compared classical Shannon information with Kolmogorov Complexity and concluded that they are compatible with respect to their ontological status. Can we come to a similar conclusion for the quantum case?

Quantum Shannon Information

As the analysis in the previous sections has shown, the only major aspect in regard to the ontological status of information that has changed in respect to the classical case is conventionality. Admittedly, Quantum Information has many other 'new' properties that classical Shannon Information doesn't have, but as explained in the cases of Quantity and Type Information, uncertainty, and the semantic/syntactic distinction, they have no effect on our already earlier made arguments. So similarly to the the classical case, Quantum Shannon Information characterizes the compressibility of a quantum information source that emits quantum tokens according to a certain probability distribution. On this view, Quantum Quantity and Type Information both remain abstracta that obey the semantic/syntactic distinction and don't suffice as a measure of uncertainty.

Regarding the degree of conventionality, Quantum Shannon Information can no longer be largely stipulated by human choice–to some degree

quantum theory fixes the experimental set up and sets a lower bound of fine graining. However, how to account for the conventionality of the partitioning still remains an open topic.

Quantum Kolmogorov Complexity

Quite analogously to Quantum Shannon Information, our analysis has shown that the main difference regarding the ontological status of QC in respect to the classical case is once again 'conventionality'. Neither the introduction of $\mathcal{QTM}$s nor the examination of qubit strings (instead of regular strings) has shifted our views about the abstractness of Quantity and Type Information, Kolmogorov Complexity as a measure of uncertainty, and the relation of semantic information with QC. Quantum theory fixes some of our initial concerns about the user dependent contingencies. However, while our fine graining is to some extent restricted, quantum theory has us to conventionally choose an orientation of our computational basis and doesn't resolve the conventional issues around coarse graining.

The Relation of the Information Measures

Lastly, a few words about the relation between Quantum Shannon Information and Quantum Kolmogorov Complexity. At the end of Chapter 6 we concluded that the classical notions of Shannon Information and Kolmogorov Complexity are compatible in the sense that neither of the two infringe the semantic/syntactic distinction nor ought to be conceived as a measure of uncertainty. Both information measures turned out to be abstract entities. In addition, both notions of information are connected through Coding Theory and each substantially depends on the contingencies of the user.

In respect to the quantum case, only the latter two aspects have changed notably. First, instead of the intuitive appealing connection that the expected Kolmogorov Complexity asymptotically equals Shannon Information, we presented a quantum version of the Brudno Theorem in section (7.4). Second, we argued that the insights of quantum theory affect the

large degree of conventionality of the classical case.

Note, whereas the above aspects are certainly different in respect to the classical case, they don't dramatically change the compatibility of the ontological statuses of the quantum information measures. Both Quantum Shannon Information and Quantum Kolmogorov Complexity, appear to be generalizations of their respective classical versions. Even though the precise connection of how they're formally related has changed in regard to the classical case, they are both still related! Additionally, while being almost completely conventional in the classical case, both are to some extent less dependent on the stipulations of the user in the quantum case; the experimental set up and the extraction of patterns (both responsible for picking out the tokens we examine), are much more restricted by having to pick out 'quantum tokens'. In the case of Quantum Kolmogorov Complexity we are then additionally confronted with stipulating a computational basis, which due to the Quantum Brudno Theorem though, doesn't appear to have a large effect on the value of QC. Lastly, the subjectivity of partitioning and coarse graining seem to remain unresolved in either case. Quite remarkably, the extent the degrees of conventionality the here compared information measures go hand in hand–in the classical case both are largely conventional, in the quantum context much less so.

Chapter 9

Results & Outlook

We started this thesis with the observation that we appear to live in an Information Age, but that the absence of a theoretical underpinning of 'information' is somewhat scandalizing. One feature of said Information Age is the application of information theoretical aspects into modern (quantum) physics. Even though physics quite successfully incorporates so called 'measures of syntactic information', each in fact having their own rigorous theoretical underpinning, we are still faced with a 'scandalizing' huge variety of claims about the ontological status of information in physics.

In this thesis we were able to partially reduce 'the scandal' by showing that the arguably two most prominent syntactic information measures–Shannon Information $H(X)$ and Kolmogorov Complexity $K(x)$–have quasi matching ontological statuses. Our analysis has shown that in the classical case, each information measure is independent of semantic information and shouldn't be regarded as a measure of uncertainty. Furthermore, $H(X)$ and $K(x)$ are related through the insights of Coding Theory and each notion of information ought to be conceived as an abstract entity. Based on a type/token distinction, we showed that concrete tokens instantiate abstract *Type Information*. However, in the classical case both information measures suffer from a lack of natural constraints that determine what the tokens actually instantiate. As a result, Shan-

non Information and Kolmogorov Complexity have to almost entirely rely on the conventional stipulations of the users.

Regarding the quantum case, we could demonstrate that our results about the semantic/syntactic distinction, uncertainty, and abstractness remain unaltered; the arguments from the classical case could virtually be copied. Due to the 'natural constraints' of quantum theory though, the degree of conventionality has to some extent decreased in respect to the classical case. The users can no longer arbitrarily stipulate any kind elementary events of a communication system or extract patterns from all kinds of 'objects' which are thought to instantiate one of the respective Type Information kinds. The choice of partitioning and coarse graining–and in the case of QC additionally the choice of a computational basis–seem to remain a substantial factor of convention though.

A few remarks. Reconsidering the claims about the ontological status of information in physics, it's neither the abstract nature of $H(X)$ nor $K(x)$ which stops us from conceiving such information measures as 'physical'. In case we don't equate 'physical' with 'material' but with 'describable by the means of physics', we might deem a lot of abstract entities (e.g., entropy or heat) as being 'physical'. Instead of the abstractness then, the crux lies in the large degree of conventionality of information in physics. Since in the classical case Shannon Information and Kolmogorov Complexity are almost completely conventional, they offer no grounds to be regarded as being part as an independent entity in the catalog of the world's furniture. Regarding the respective quantum notions however, the case might look different as the conventional choices by the users are in part replaced with the 'natural constraints' of quantum theory. To what extent such a replacement suffices to point out either notion of quantum information as 'independent of us' remains an open question and ought to be the direction of further research.

Chapter 10

Appendix

10.1 Letter frequencies

letter	frq [%]	letter	frq [%]	letter	frq [%]	letter	frq [%]
e	12.02	y	2.11	e	16.93	b	1.96
t	9.10	w	2.09	n	10.53	w	1.78
a	8.12	g	2.03	i	8.02	f	1.49
o	7.68	p	1.82	r	6.89	k	1.32
i	7.31	b	1.49	s	6.42	z	1.21
n	6.95	v	1.11	t	5.79	v	0.84
s	6.28	k	0.69	a	5.58	p	0.67
r	6.02	x	0.17	h	4.98	ü	0.65
h	5.92	q	0.11	d	4.98	ä	0.54
d	4.32	j	0.10	u	3.83	ß	0.37
l	3.98	z	0.07	l	3.60	ö	0.30
u	2.88	ä	n.d.	c	3.16	j	0.24
c	2.71	ö	n.d.	g	3.02	y	0.05
m	2.61	ü	n.d.	m	2.55	x	0.05
f	2.30	ß	n.d.	o	2.24	q	0.02

Table 10.1: Letter frequencies in English and German.

10.2 H(X) from L(X)

Let us now derive the Shannon information measure from considerations of Coding theory according to [Lyre, 1998]. We want to minimize the

mean codeword length $L(X)$ (4.3) over all idealized codes C, such that

$$H(P) = \min_{C} \sum_{i} p_i l_i. \tag{10.1}$$

Hence, we identify (10.1) as the function f, which ought to be minimized

$$f(l_i) = \sum_{i} p_i l_i =^{!} \min. \tag{10.2}$$

Moreover, we obtain our constraint g due to the Kraft inequality (4.6)

$$g(l_i) = \sum_{i} 2^{-l_i} - 1 = 0, \tag{10.3}$$

such that, with the help of the method of Lagrange multipliers we get

$$\mathcal{L}(l_i, \lambda) = f(l_i) + \lambda g(l_i). \tag{10.4}$$

Deriving to l_i and λ, yields

$$\frac{\partial \mathcal{L}(l_i, \lambda)}{\partial l_i} = p_i - \lambda \cdot \log 2 \cdot 2^{-l_i} = 0 \tag{10.5}$$

and

$$\frac{\partial \mathcal{L}(l_i, \lambda)}{\partial \lambda} = g(l_i) = 0. \tag{10.6}$$

Adding (10.5) and (10.6) yields

$$\sum_{i} p_i - \lambda \log 2 \cdot \sum_{i} 2^{-l_i} = 0. \tag{10.7}$$

With the side conditions (10.5) and (10.6) we solve for

$$\lambda = \frac{1}{\log 2} \tag{10.8}$$

and eventually get

$$l_i = -\log p_i \tag{10.9}$$

as the value to minimize the mean code word length $L(X)$ (10.1). Plugging in our result, we indeed obtain

$$\min_C \sum_i p_i l_i = -\sum_i p_i \log p_i. \tag{10.10}$$

10.3 Shannon's axiomatic derivation

In his original paper [Shannon, 1948], Shannon presented the following postulates for arriving at his information measure $H(X)$.

> "Suppose we have a set of possible events whose probabilities of occurrence are $p_1, p_2, ..., p_n$. These probabilities are known but that is all we know concerning which event will occur. Can we find a measure of how much "choice" is involved in the selection of the event or of how uncertain we are of the outcome? If there is such a measure, say $H(p_1, p_2, ..., p_n)$, it is reasonable to require of it the following properties:
>
> 1. H should be continuous in the p_i.
>
> 2. If all the p_i are equal, $p_i = 1/n$, then H should be a monotonic increasing function of n. With equally likely events there is more choice, or uncertainty, when there are more possible events.
>
> 3. If a choice be broken down into two successive choices, the original H should be the weighted sum of the individual values of H. The meaning of this is illustrated in Fig. (10.1).

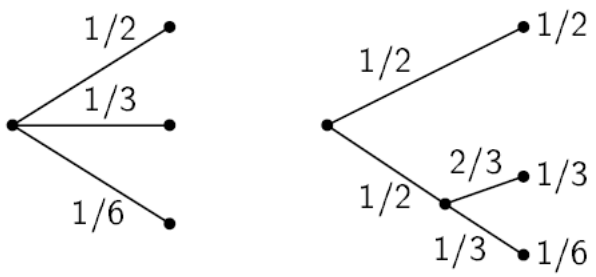

Figure 10.1: Decomposition of choice from three possibilities.

At the left we have three possibilities $p_1 = \frac{1}{2}$, $p_2 = \frac{1}{3}$, $p_3 = \frac{1}{6}$. On the right we first choose between two possibilities each with probability 12 , and if the second occurs make another choice with probabilities 23 , 13 . The final results have the same probabilities as before. We require, in this special case, that

$$H(\frac{1}{2}, \frac{1}{3}, \frac{1}{6}) = H(\frac{1}{2}, \frac{1}{2}) + \frac{1}{2}H(\frac{2}{3}, \frac{1}{3})$$

The coefficient $\frac{1}{2}$ is the weighing factor introduced because this second choice only occurs half the time." [Shannon, 1948, p. 392]

Whereas the first two postulates are formulated as exact rules, the third requirement is only delivered in form of an example. In absence of a completely rigorous proof, Shannon stated that the assumptions for proving the uniqueness of $H(X)$ wouldn't be necessary for his present theory. He concludes that

"[t]he real justification of these definitions, however, will reside in their implications.", (ibid. p. 393).

Lastly Shannon then presented his famous uniqueness theorem, stating that $H = -K \sum_{i=1}^{n} p_i \log p_i$ is the only H satisfying the above assumptions.

Other axiomatic derivations

The not yet exactly formulated third requirement, soon led to many sets of postulates (in fact axiomatic systems), all having Shannon's information measure as unique solution [Aczeel and Daroczy, 1975]. Faddeev could demonstrate that

"If the expression $H_n(p_1, ..., p_n)$ for $p_i \geq 0$, $\sum_i p_i = 1$and $n \geq 2$ satisfies the conditions:

1. $H_2(p, 1-p)$ is a continuous positive function of p;
2. for all n, $H_n(p_1, ..., p_n)$ is a symmetrical, (i.e. permutation invariant) function of $p_1, ..., p_n$;
3. for all $n \geq 2$,

$$H_n(p_1, ..., p_n) = H_{n-1}(p_1+p_2, p_3, ..., p_n)+(p_1+p_2)H_2\left(\frac{p_1}{p_1+p_2}, \frac{p_2}{p_1+p_2}\right);$$

then $H_n$has the form

$$H_n = -K\sum_{i=1}^{n} p_i \log p_i$$

for some positive constant K." [Hilgevoord and Uffink, 1991]

However, as pointed out in [Uffink, 1991, §1.6.3], while such axiomatic systems uniquely derive (2.10), it doesn't follow that they are automatically free from further concerns. Faddeev's axiomatic system for instance, is not free of such problems, too. While the first two requirements are quite straight forward and are pretty much in accordance with Shannon's formulation (or can e.g., otherwise easily 'read off' from fig. 2.3) the third axiom needs some further scrutiny. Faddeev's third axiom, also known as *recursion requirement*, generalizes Shannon's idea that $H(X)$ should be the weighted sum of the individual values of $H(X)$, once a choice is broken down into two successive choices. According to Uffink, the problems with Faddeev's third axiom are i) conventional aspects, ii) divergence when the number of possible outcomes is unbounded, and iii) when Shannon's information measure is generalized to continuous probability functions.

As pointed out earlier though, we can find other sets of postulates, some of which don't rely on the recursion requirement. However, these sets of postulates don't characterize Shannon's *relative* information measure in a unique way. The notion of a relative information measure is on the other hand crucial to account for continuous probability density functions. Whether one wishes to derive Shannon's information measure for the discrete *and* the continuous case, depends to some extent on one's requirements for the postulate system. For sake of completeness it is therefore instructive to take a closer look at iii) in the following subsection.

The relative information measure for continuous probability distributions

It is often argued that Shannon's information measure doesn't have a natural extension to the continuous case. When (2.10) is generalized to a continuous probability density function $p(x),\ x\ \in \mathbb{R}$, we obtain

$$H(P) = -\int p(x) \log p(x)\ dx \tag{10.11}$$

as an analogue (which was already suggested by Shannon himself). Yet the continuous case (10.11) has a couple of differing properties from its discrete counterpart (2.10). First of all, expression (10.11) may take negative values. Moreover, Uffink demonstrates [Uffink, 1991] that (10.11) depends on the (arbitrary) labeling of the outcomes.

In order to circumvent the above mentioned problems by the transition of the discrete case (2.10) of the *absolute* information measure to the continuous case, Uffink [Uffink, 1991], [Uffink, 1995] suggests to introduce a 'background measure' μ. With the positive weights $\mu(x_i)$ determined by the background measure, we obtain the *relative* information measure

$$H(P, \mu) = -\sum p_i \log \frac{p_i}{\mu_i}. \tag{10.12}$$

First note, that by choosing the counting measure (i.e. if $\forall i\ :\ \ \mu(x_i) =$

1), the relative information measure (10.12) becomes equivalent to the absolute information measure (2.10) again. For a continuous probability function P, we can rewrite the relative information measure with respect to a background measure μas

$$H(P,\mu) = -\int \frac{\partial P}{\partial \mu}(x) \log \frac{\partial P}{\partial \mu}(x) \mathrm{d}\mu(x) \tag{10.13}$$

By now choosing μ to be the Lebesgue measure λ, expression (10.13) reduces to

$$H(P,\lambda) = -\int p(x) \log p(x)\ dx, \tag{10.14}$$

hence effectively being equivalent with (10.11) [Uffink, 1995].

In order to define Shannon Information for continuous probability distributions, we seemingly should replace the concept of absolute information with that of relative information. Depending on how desirable we deem the successful derivation of the continuous case, we therefore might have to reject Faddeev's third axiom (which can only account for the absolute case). Overall the choice of an axiomatic system to derive $H(X)$ remains a delicate matter.

10.4 Approaches to Quantum Complexity

- On of the first approaches is due to Svozil [Svozil, 1996, §5]. Svozil defines the algorithmic complexity

$$H(s) = \min_{C(p)=s} l(p), \tag{10.15}$$

of a vector $s \in \mathcal{H}$ in some Hilbert space $\mathcal{H}$ as the length of the shortest program p running on a quantum computer C as reference machine. For reasons of maintaining the convergence of the Kraft inequality, Svozils approach is restricted to purely *classical* prefix-free programs as input; in case where quibits were allowed, the Kraft inequality would diverge (that said, the output is allowed to be quantum though). However, because the number of classical binary strings is finite, this definition has the disadvantage to not

be able to account for the infinitely many states of $s \in \mathcal{H}$.

- In [Vitanyi, 2001], Vitanyi provided a similar definition (classical input, yet quantum output), but circumventing Svozil's problems by allowing errors and non-perfect results. This way, Vitany circumvents the disadvantages of Svozil's account and now being able to account for the infinitely many states $s \in \mathcal{H}$ within a certain margin of error. On this view, the output $Q(p)$ of the quantum computer Q with input p, doesn't have to be exactly equivalent with the analyzed state $|x\rangle$. That's why Vitanyi's defintion of the quantum algorithmic information measure

$$K\left(|x\rangle\right) = \min\left\{l(p) + \left\lceil -\log \|\langle z|x\rangle\|^2 \right\rceil : \; Q(p) = |z\rangle\right\} \tag{10.16}$$

  introduces the 'penalty term' $\left\lceil -\log \|\langle z|x\rangle\|^2 \right\rceil$. The former expression is based on the *fidelity* $\|\langle x|z\rangle\|^2$which measures how 'close' the vectors $|x\rangle$and $|z\rangle$are. If $Q(p)$ and $|z\rangle$ differ too much, then the penalty term becomes so large, that the shortest program has to be found with another argument than p.

- Gacs [Gacs, 2001] on the other hand, has introduced an approach based on universal probability (4.10) and seems to be influenced by Levin's concept of a universal semicomputable (semi)measures.[1] Gacs showed that taking a universal semicomputable density matrix μ (basically the quantum analog to a classical probability distribution) can be used to derive a form of quantum complexity as the negative logarithm of μ. It is a striking feature of his two information measures (depending on the order of taking the logarithm)

$$\underline{H} = -\log \langle\psi|\, \mu \,|\psi\rangle \tag{10.17}$$

$$\overline{H} = -\langle\psi|\, (\log \mu) \,|\psi\rangle\,, \tag{10.18}$$

[1]Levin's Coding theorem is about such 'semimeasures', so that 'probability distributions' p defined on strings may sum to less than one$\sum_{x\in\{0,1\}} p(x) \leq 1$. We call a semimeasure 'semicomputable' if there's a monotonically increasing, computable sequence of functions converging to it (cf. [Mueller, 2007]).

that they come without reference to any model of neither classical nor quantum computation. Overall, Gacs then establishes the connection to quantum complexity by demonstrating that (10.17) and (10.18) are the lower and upper bounds of Vitanyi's $K(|x\rangle)$ (10.16) respectively.

- The first purely quantum based approach (i.e. when considering a quantum reference machine and quantum in- and output strings) was developed by Berthiaume, van Dam and Laplante[Berthiaume et al., 2001]. Their defintion reads as

$$QC^{\alpha}(|\psi\rangle) = \min\left\{l(|\varphi\rangle) \mid \langle\psi| U(|\varphi\rangle)|\psi\rangle \geq \alpha\right\}, \tag{10.19}$$

  where U is a universal quantum computer and the complexity of $|\psi\rangle$ is defined as the shortest *quantum* input $|\varphi\rangle$ which produces $|\psi\rangle$ to some *fidelity* greater than α. If $\alpha = 1$, then $U(|\varphi\rangle)$ is equal to $|\psi\rangle$, whereas for $\alpha < 1$ a certain degree of inaccuracy is allowed for. A slightly modified version of this approach is the basis for our presentation in Chapter 7.

- Yet another approach is due to by Mora and Briegel [Mora and Briegel, 2004] (see also [Mora and Briegel, 2005]), where they relate the complexity to the shortest classical description of some quantum curcuit $\mathcal{C}$ that prepares the state. A quantum circuit can be regarded as a sequence of elementary operations characterizing a quantum state, where the complexity of a state refers to the circuit itself. Once a complete gate basis B and a code Ω are fixed, and a circuit $\mathcal{C}^{B,\varepsilon}$that prepares $|\varphi\rangle$with precision ε,[2] then

$$K_{\mathbf{Net}}^{\Omega,B,\varepsilon}(|\varphi\rangle) = \min_{\mathcal{C}^{B,\varepsilon}\in\widetilde{\mathcal{C}}^{B,\varepsilon}} K_{\mathbf{Net}}^{\Omega,B,\varepsilon}(|\varphi\rangle) \tag{10.20}$$

  defines the quantum algorithmic complexity, according to this approach.

[2]Determining the precision here, can be understood as referring to two states $|\varphi\rangle$and $|\psi\rangle$ as ε-*distinguishable* if $\|\langle\varphi|\psi\rangle\|^2 \leq 1-\varepsilon$

Bibliography

[Aczeel and Daroczy, 1975] Aczeel, J. and Daroczy, Z. (1975). *On Measures of Information and Their Characterizations*. Academic Press.

[Adriaans, 2013] Adriaans, P. (2013). Information. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, fall 2013 edition.

[Aguirre et al., 2015] Aguirre, A., Foster, B., and Merali, Z., editors (2015). *If From Bit or Bit From It?* Springer International Publishing.

[Barker-Plummer, 2016] Barker-Plummer, D. (2016). Turing machines. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, winter 2016 edition.

[Benatti, 2009] Benatti, F. (2009). *Dynamics, Information and Complexity in Quantum Systems*. Springer Netherlands.

[Benatti et al., 2006] Benatti, F., Krueger, T., Mueller, M., Siegmund-Schultze, R., and Szkola, A. (2006). Entropy and Quantum Kolmogorov Complexity: A Quantum Brudno's Theorem. *Communications in Mathematical Physics*, 265:437–461.

[Bennett, 1982] Bennett, C. H. (1982). The thermodynamics of computation - a review. *International Journal of Theoretical Physics*, 21(12):905–940.

[Bernstein and Vazirani, 1997] Bernstein, E. and Vazirani, U. (1997). Quantum complexity theory. *SIAM Journal on Computing*, 26(5):1411–1473.

[Berthiaume et al., 2001] Berthiaume, A., van Dam, W., and Laplante, S. (2001). Quantum kolmogorov complexity. *Journal of Computer and System Sciences*, 63(2):201–221.

[Brudno, 1982] Brudno, A. (1982). Entropy and the complexity of the trajectories of a dynamic system. *Trudy Moskovskogo Matematicheskogo Obshchestva*, 44:124–149.

[Brukner and Zeilinger, 2001] Brukner, C. and Zeilinger, A. (2001). Conceptual inadequacy of the shannon information in quantum measurements. *Physical Review A, Volume 63*.

[Bub, 2007] Bub, J. (2007). Quantum information and computation. In Butterfield, J. and Earman, J., editors, *Philosophy of Physics Part A*, pages 555–660.

[Bub, 2012] Bub, J. (2012). Is information the key? In *Analysis and Interpretation in the Exact Sciences*, pages 219–233. Springer.

[Bub, 2016] Bub, J. (2016). *Bananaworld*. Oxford University Press.

[Capurro, 1978] Capurro, R. (1978). *Information*. PhD thesis.

[Capurro, 2009] Capurro, R. (2009). Past, present, and future of the concept of information. *tripleC: Communication, Capitalism & Critique. Open Access Journal for a Global Sustainable Information Society*, 7(2):125–141.

[Capurro and Hjorland, 2003] Capurro, R. and Hjorland, B. (2003). The concept of information. *Annual Review of Information Science and Technology*, 37(1):343–411.

[Chaitin, 1966] Chaitin, G. (1966). On the length of programs for computing finite binary sequences. *Journal of the ACM*, pages 547–569.

[Chaitin, 1975] Chaitin, G. (1975). Randomness and mathematical proof. *Scientific American 232*, (5):47–52.

[Chaitin, 1982] Chaitin, G. (1982). Goedel's theorem and information. *International Journal of Theoretical Physics*, 21(12):941–954.

[Chaitin, 1986] Chaitin, G. (1986). Randomness and goedel's theorem. *Mondes en Developpement*, (54-55):125–128.

[Copeland, 2015] Copeland, B. J. (2015). The church-turing thesis. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, summer 2015 edition.

[Cover and Thomas, 2006] Cover, T. M. and Thomas, J. A. (New Jersey 2006). *Elements of Information Theory*. John Wiley & Sons, 2 edition.

[Denbigh and Denbigh, 1985] Denbigh, K. G. and Denbigh, J. S. (1985). *Entropy in relation to incomplete Knowledge*. Cambridge: Cambridge University Press.

[Desurvire, 2009] Desurvire, E. (2009). *Classical and quantum information theory: an introduction for the telecom scientist*. Cambridge University Press.

[Deutsch, 1985] Deutsch, D. (1985). Quantum theory, the church-turing principle and the universal quantum computer. In *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, volume 400, pages 97–117.

[Devine, 2009] Devine, S. (2009). The insights of algorithmic entropy. *entropy*, (11):85–110.

[Duwell, 2003] Duwell, A. (2003). Quantum information does not exist. *Studies in History and Philosophy of Science Part B*, 34(3):479–499.

[Duwell, 2008] Duwell, A. (2008). Quantum information does exist. *Studies in History and Philosophy of Science Part B*, 39(1):195–216.

[Duwell, 2017] Duwell, A. (2017). Representation, interpretation, and theories of information. In Olimpia Lombardi, Sebastian Fortin, F. H. C. L., editor, *What Is Quantum Information?* Cambridge University Press.

[Feynman, 1982] Feynman, R. P. (1982). Simulating physics with computers. *International journal of theoretical physics*, 21(6):467–488.

[Fisher, 1925] Fisher, R. (1925). Theory of statistical estimation. *Mathematical Proceedings of the Cambridge Philosophical Society*, 5(22):700–725.

[Frigg and Nguyen, 2016] Frigg, R. and Nguyen, J. (2016). Scientific representation. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, winter 2016 edition.

[Gacs, 2001] Gacs, P. (2001). Quantum algorithmic entropy. *Journal of Physics A: Mathematical and General*, 34(35):6859.

[Galavotti, 2005] Galavotti, M. C. (2005). *A Philosophical Introduction to Probability*. CSLI Publications.

[Goguen, 1997] Goguen, J. A. (1997). Towards a social, ethical theory of information. In Geoffrey C. Bowker, Susan Leigh Star, W. T. L. G., editor, *Social Science, Technical Systems and Cooperative Work: Beyond the Great Divide*, pages 27–56. Erlbaum.

[Grimus, 2011] Grimus, W. (2011). On the 100th anniversary of the sackur-tetrode equation. *ArXiv e-prints:1112.3748*.

[Gruenwald and Vitanyi, 2004] Gruenwald, P. and Vitanyi, P. M. B. (2004). Shannon information and kolmogorov complexity. *CoRR*, cs.IT/0410002.

[Gruenwald and Vitanyi, 2008] Gruenwald, P. D. and Vitanyi, P. (2008). Algorithmic information theory. *arXiv preprint arXiv:0809.2754*.

[Harremoes and Topsoe, 2008] Harremoes, P. and Topsoe, F. (2008). The quantitative theory of information. In Adriaans, P. and van Benthem, J., editors, *Philosophy of Information*, pages 171–216. Elsevier.

[Hartley, 1928] Hartley, R. (1928). Transmission of information. *Bell System Technical Journal*, (7):535–563.

[Hilgevoord and Uffink, 1991] Hilgevoord, J. and Uffink, J. (1991). Uncertainty in prediction and in inference. *Foundations of Physics*, 21(3):323–341.

[John E. Hopcroft and Ullman, 2001] John E. Hopcroft, R. M. and Ullman, J. D. (2001). *Introduction to Automata Theory, Languages and Computation*. Addison-Wesley, 2 edition.

[Jozsa, 1994] Jozsa, R. (1994). Fidelity for mixed quantum states. *Journal of Modern Optics*, 41(12):2315–2323.

[Jozsa, 2004] Jozsa, R. (2004). Illustrating the concept of quantum information. *IBM Journal of Research and Development*, page 0305114.

[Kolmogorov, 1933] Kolmogorov, A. N. (1933). *Grundbegriffe der Wahrscheinlichkeitsrechnung*. Berlin: Springer.

[Kolmogorov, 1965] Kolmogorov, A. N. (1965). Three approaches to the quantitative definition of information. *Problemy Peredachi Informatsii*, 1(1):3–11.

[Landauer, 1991] Landauer, R. (1991). Information is physical. *Physics Today*, 44:23–29.

[Landauer, 1996] Landauer, R. (1996). The physical nature of information. *Physics Letters A*, 217(4):188 – 193.

[Landauer, 1999] Landauer, R. (1999). Information is a Physical Entity. *Physica A Statistical Mechanics and its Applications*, 263:63–67.

[Li and Vitanyi, 2008] Li, M. and Vitanyi, P. (2008). *An Introduction to Kolmogorov Complexity and Its Applications*. Springer: New York, 3 edition.

[Li and Vitanyi, 1992] Li, M. and Vitanyi, P. M. B. (1992). *Philosophical issues in Kolmogorov complexity*, pages 1–15. Springer Berlin Heidelberg, Berlin, Heidelberg.

[Lloyd, 2009] Lloyd, S. (2009). Quantum information science (lecture notes).

[Lyre, 1998] Lyre, H. (1998). *Quantentheorie der Information*. PhD thesis.

[MacLeod and Rubenstein, 2017] MacLeod, M. C. and Rubenstein, E. M. (2017). Universals. In *Internet Encyclopedia of Philosophy*.

[Maudlin, 2007] Maudlin, T. (2007). *The Metaphysics Within Physics*. Oxford University Press.

[Maurin, 2016] Maurin, A.-S. (2016). Tropes. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, winter 2016 edition.

[Miszczak et al., 2008] Miszczak, J. A., Puchala, Z., Horodecki, P., Uhlmann, A., and Zyczkowski, K. (2008). Sub–and super–fidelity as bounds for quantum fidelity. *arXiv preprint arXiv:0805.2037*.

[Mora et al., 2006] Mora, C., Briegel, H., and Kraus, B. (2006). Quantum kolmogorov complexity and its applications. *eprint arXiv:quant-ph/0610109*.

[Mora and Briegel, 2004] Mora, C. and Briegel, H. J. (2004). Algorithmic complexity of quantum states. *eprint arXiv:quant-ph:0412172*.

[Mora and Briegel, 2005] Mora, C. E. and Briegel, H. J. (2005). Algorithmic Complexity and Entanglement of Quantum States. *Physical Review Letters*, 95(20):200503.

[Mueller, 2007] Mueller, M. (2007). *Quantum Kolmogorov Complexity and the Quantum Turing Machine*. PhD thesis.

[Mueller and Rogers, 2008] Mueller, M. and Rogers, C. (2008). Quantum Bit Strings and Prefix-Free Hilbert Spaces. *ArXiv e-prints:0804.0022*.

[Nielsen and Chuang, 2000] Nielsen, M. A. and Chuang, I. L. (2000). *Quantum Computation and Quantum Information*. Cambridge University Press.

[Nyquist, 1924] Nyquist, H. (1924). Certain factors affecting telegraph speed. *Bell System Technical Journal*, pages 324–346.

[Rogers and Vedral, 2008] Rogers, C. C. and Vedral, V. (2008). The second quantized quantum turing machine and kolmogorov complexity. *Modern Physics Letters B*, Vol.22(No.12):1203–1210.

[Schumacher, 1995] Schumacher, B. (1995). Quantum coding. *Phys. Rev. A*, 51:2738–2747.

[Shannon, 1948] Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, 27:379–423, 623–656.

[Shannon, 1993] Shannon, C. E. (1993). *The Lattice Theory of Information*, pages 180–183. Wiley Interscience.

[Solomonoff, 1964] Solomonoff, R. J. (1964). A formal theory of inductive inference part i & ii. *Information and Control*, (7):1–22, 224–254.

[Solomonoff, 1997] Solomonoff, R. J. (1997). The discovery of algorithmic probability. *Journal of Computer and System Sciences*, 55(1):73–88.

[Svozil, 1996] Svozil, K. (1996). Quantum algorithmic information theory. *Journal of Universal Computer Science*, 2:311–346.

[Timpson, 2004] Timpson, C. G. (2004). *Quantum Information Theory and The Foundations of Quantum Mechanics*. PhD thesis.

[Timpson, 2006] Timpson, C. G. (2006). Philosophical aspects of quantum information theory. *arXiv preprint quant-ph/0611187*.

[Timpson, 2010] Timpson, C. G. (2010). Information, immaterialism, instrumentalism: Old and new in quantum information. *Philosophy of quantum information and entanglement*, pages 208–227.

[Timpson, 2013] Timpson, C. G. (2013). *Quantum Information Theory & the Foundations of Quantum Mechanics*. Oxford: Clarendon Press.

[Tribus and McIrvine, 1971] Tribus, M. and McIrvine, E. (1971). Energy and information. *Scientific American*, (224):178–184.

[Uffink, 1991] Uffink, J. (1991). *Measures of Uncertainty and the Uncertainty Principle*. PhD thesis.

[Uffink, 1995] Uffink, J. (1995). Can the maximum entropy principle be explained as a consistency requirement? *Studies in History and Philosophy of Science Part B*, 26(3):223–261.

[Uhlmann, 1976] Uhlmann, A. (1976). The transition probability in the state space of a*-algebra. *Reports on Mathematical Physics*, 9(2):273 – 279.

[van Lunteren, 2016] van Lunteren, F. (2016). Clocks to computers: A machine-based "big picture" of the history of modern science. *Isis*, 107(4):762–776.

[Vitanyi, 2001] Vitanyi, P. M. (2001). Quantum kolmogorov complexity based on classical descriptions. *IEEE Transactions on Information Theory*, 47(6):2464–2479.

[Vitanyi, 2010] Vitanyi, P. M. (2010). Ray solomonoff, founding father of algorithmic information theory. *Algorithms*, 3(3):260–264.

[Vitanyi, 2012] Vitanyi, P. M. B. (2012). Conditional kolmogorov complexity and universal probability. *CoRR*, abs/1206.0983.

[Wheeler, 1989] Wheeler, J. A. (1989). Information, physics, quantum: The search for links. In Zurek, W. H., editor, *Complexity, Entropy and the Physics of Information*, volume 8, pages 3–28. Addison-Wesley.

[Zeilinger, 2004] Zeilinger, A. (2004). *Why the quantum? It from bit? A participatory universe? Three far-reaching challenges from John Archibald Wheeler and their relation to experiment*, pages 201–220.

[Zurek, 1989] Zurek, W. H. (1989). Algorithmic randomness and physical entropy. *Physical Review A*, 40(8):4731–4751.

[Zvonkin and Levin, 1970] Zvonkin, A. K. and Levin, L. A. (1970). The complexity of finite objects and the development of the concepts of information and randomness by means of the theory of algorithms. *Russian Mathematical Surveys*, 25(6):83.