2017

Measures of Information for Information Acquisition Optimization

Zhenglin Wei
Lehigh University

Measures of Information for Information Acquisition Optimization

by

Zhenglin Wei

Presented to the Graduate and Research Committee
of Lehigh University
in Candidacy for the Degree of
Master of Science
in
Industrial and Systems Engineering

Lehigh University
May 2017

This thesis is accepted and approved in partial fulfillment of the requirements for the Master of Science.

____
Date

____
Dr. Eugene Perevalov
Thesis Advisor

____
Dr. Tamás Terlaky
Chairperson of Department

Acknowledgements

First of all, I would like to express my gratitude to my advisor, Professor Eugene Perevalov, to whom I am deeply grateful for his encouragement, patience, and guidance during the research. The idea of his research is so great that it not only involves the traditional information theory but also has some connections with cognitive science and artificial intelligence. Without his support, I would never touch the field of Information Theory and would never have had the possibility to complete this effort.

Also, I want to genuinely thank my parents who always support me to pursue my dream through all the periods of my life. Mom and Dad, I am really lucky to have you two. Without your help, I would never have had the opportunity to learn outside of my country and open my eyes to the global world.

Finally, I would like to thank all of my family members and friends for their help and concern from all over the world.

Contents

List of Tables

List of Figures

Abstract

The objective of this study is to find suitable methods of information measurement and characterizations to facilitate research on information acquisition optimization. Specifically, this study is to support an approach, which has been acquired in past periods of the research, that can be interpreted as a theory of information exchange between the decision maker and information source(s). It can also be said that the developed approach complements the classical Information Theory. The classical Information Theory describes transmission of information over some channel, regardless of its content. The proposed approach deals the first and last link of the full information chain: extracting information from a source and using it to obtain the best possible decision. In this approach, a quantitative framework describing *information sources*, the process of information exchange between the source(s) and the agent, and the relevance of the obtained information for the given problem is developed. The purpose of this thesis is to study and summarize the existing methods of measuring information and evaluate their suitability for the purposes of the proposed approach.

Chapter 1

Introduction

1.1 Current Research: Information Acquisition

The objective of the overall research is to develop a theory of information acquisition for quantitative decision making. This work can also be looked upon as an attempt to make Information Theory methods useful for optimization and decision making under uncertainty. If the information is to be used to solve problems, and the goal is to maximize the solution quality (e.g. by minimizing the properly defined loss) by making use of available information sources, then one needs to know what specific information is to be requested from a source so that, on one hand, the source would be able to fulfill the request accurately and, on the other hand, the information obtained would have a large impact on the solution quality for the specific problem at hand. Some preliminary results can be found in [39, 37, 38, 40]. (Please note : our research is still ongoing. Many materials are being revised or to be improved. The results in different stages of the research in the future may be similar to the following sections but changes and differences may occur if necessary.)

We believe that progress on this particular front has been made especially important by the recent advance of the so called "Big Data", i.e. the newly acquired ability to record, store and manipulate large amounts of data of various flavors. The availability of such large amounts of information holds a promise, but also presents a challenge. The promise is related to the "bigness" of the data: it is likely to have a lot to do with almost any problem because of the large amount of data (since everything is interrelated). The challenge, not surprisingly, is directly related to the very same "bigness". Since the amounts of the data is so large and potentially related to almost everything, it is difficult to find and extract useful bits of it for a given problem. Therefore, in order to be able to use large amounts of data, one needs to learn how to extract useful information.

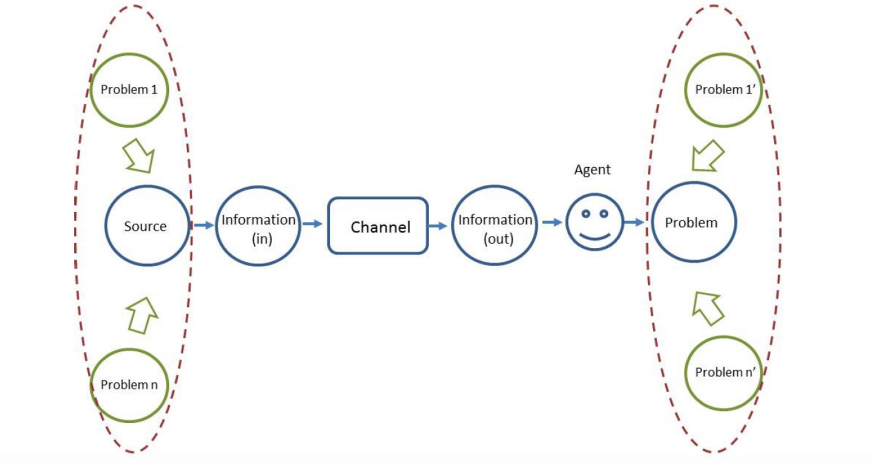
1.1.1 Comparison with Classical Information Theory

The above considerations naturally lead one to the need of a theory of information flow. Such theory seems to exist - the Information Theory. However, if one opens a textbook on the classical Information Theory, a mild surprise might ensue in that a definition of information itself is typically lacking. If one reads carefully, it is possible to see that information is understood as a collection of symbols from a certain alphabet sets that is assumed to be given at the outset and is in need of transmission to a different point in space – as quickly as possible and hopefully without errors – even though the transmission channel might not be error-free. This theory gives little about information extraction. The question is why this is so and what else we need to know to extract accurate information from a variety of sources. The existing information theory focuses on the image portion of the information, ignoring what its image represents. In other words, it deals with the external form of information. If one just needs to transmit and compress, it is sufficient. If information needs to be extracted from a source, it is not very useful. One needs to study the essence of information: that what forms it. The underlying reason why information needs to be exchanged is the existence of a distributed (relatively isolated) nature of human activity. Because of the latter, various agents can benefit from the experience of other agents, which are communicated as information. Encoding in the various symbols is the experience gained by the agent in the process of solving certain problems. This experience is first reflected in their knowledge, which we believe should be seen as the essence of information.

Classical Information Theory solves the problem of maximizing the quantity of information that can be reliably transmitted over the given (imperfect) channel, without dealing with the questions of where the information comes from and what it's going to be used for. In the Classical Information Theory, one abstracts from the source and then the requester (also the problem) and takes information as given. Thus one also abstracts from the content of information(since the latter is determined by the problem). In the developed theory, it abstract from the broader context of the problem: why this problem is being solved and what for. And also abstract from any details of the activity that led to the formation of the source's knowledge. It can be said that the developed methodology complements the classical Information Theory in that it deals in the context of quantitative decision making at least with the first and last link of the full "information chain": extracting it from the source and using it to obtain the best possible decision. The classical Information Theory describes the middle link of that chain in case a transmission of the information obtained from the source over some channel is involved. The middle link just happens to be largely independent of the end links and can be treated separately, while the end links are rather closely connected and therefore have to be treated together.

Here, three figures shows three different context of problem solving by agents. The classical Information Theory appears to be an independent part of the extended theory in that it can be considered separately with no loss of description accuracy

Figure 1.1: The Broad Context of Problem Solving by Agents



The form of an image that used in classical IT is a pre-specified set of symbols in the alphabet. Without changing the nature of the image, the particular alphabet itself is obviously not important. In this sense, people can use different alphabets and mappings from the first to the second. In particular, you can select any "standard" alphabet and map each letter. The simplest alphabet that is usually used as a standard is a binary string with only two symbols: 0 and 1. If a person insists on using a binary alphabet, he can concentrate on the essence of the image and

Figure 1.2: Classic Information Theory

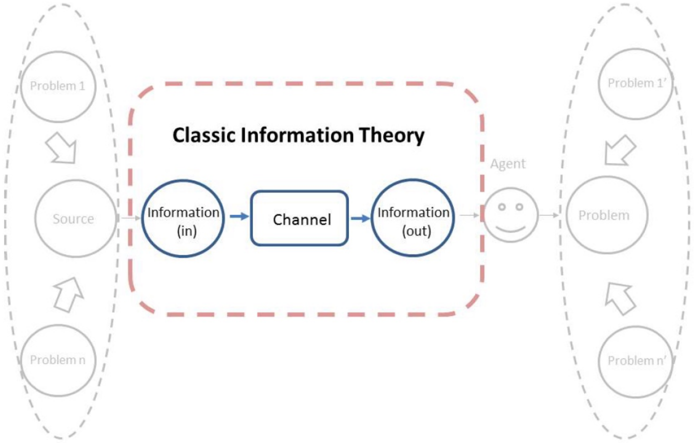
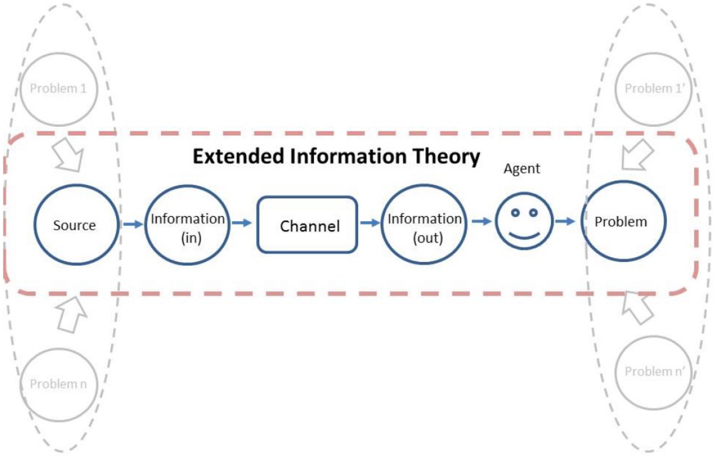


Figure 1.3: Extended Information Theory



appear it in the clearest way on the surface in binary form. This can be an intuitive study of the quantitative analysis of this essence.

If one can learn the statistical properties of the original image, regardless of the alphabet, one can use the encoding of the original image to map to the standard alphabet. The knowledge of the statistical properties can then be used to remove the redundancy from the image, that is, to minimize the expected length of the mapped image by using the Kraft inequality. The length of the best-coded image will then represent the true number of images (as opposed to the visible number in the original alphabet). Once the basic quantity is determined, the image dynamics can be extensively quantitatively studied. Especially its transmission on imperfect channels. It is possible to transmit error-free data through imperfect channels, which is one of Shannon's main results. The reason is that the imperfect channel will affect the transmission of the symbol that is the appearance of the image. Introducing planned redundancy in your code ensures that changes do not affect essence.

On the surface, the information exists in the form of symbols. The theme of this level is the subject of classical information theory. After this changing surface, there is always some knowledge. This knowledge is the crystallization of human activity. This is the essence of information at a deeper level. Essence should appear on the surface. Knowledge is exchanged by an agent, and the exchange takes place in symbolic form. The knowledge comes to the surface through the information source satisfying the agent's specific information request. We can call them the answer to the question

Finally, information only becomes actualized when it gets used to solve a problem. In this sphere, the surface, the essence and the appearance of the essence come together. It's in this sphere that the usefulness of information can be measured. If a quantitative theory is our goal we will need to tie everything to this sphere.

1.1.2 General Description of the Research and Approach

In classical Information Theory, the main question is two-fold: what is the maximum (theoretical) speed of accurate transmission for the given channel and how that speed can be (practically) achieved. The first part of the main question is addressed by calculating the channel capacity and the second part of the main question is addressed by designing appropriate codes for input symbols. The main question being addressed in the proposed approach is also two-fold: what is the maximum decision quality (for the given problem) that can be achieved by using the available information source(s) and what is the practical way of achieving that quality. The first part of this question is addressed by computing the loss efficient frontier for the given problem and source specific function, and the second part is addressed by designing appropriate questions (that lie on the efficient frontier) as means of extracting information from the source(s) optimally with respect to the decision making problem being solved.

Often, information that such sources possess fails to be taken advantage of due to its perceived and factual imprecision and to the lack of a methodology that allows for this in a controlled and regular fashion. Moreover, the above-mentioned inability to extract additional pieces of useful information often results in decisions being made simply based on decision makers' intuition and qualitative judgement because of the perceived imprecision of the available probability measure. In this research, we initiate development of a unified theoretical framework for optimal information acquisition in general purpose decision making problems including those with large and complex feasible regions to address such a situation.

The approach begins with the assumption that one or several information sources are available that are capable of providing potentially various (i.e. qualitatively different) "bits" of additional information on top of what's already contained in the initial probability measure. The assumption of having available such "multi purpose" information sources is made to describe primarily human experts that possess a certain "picture" of the way the investigated system will likely develop in the future and capable of internally "processing" that picture to answer specific questions concerning possible future outcomes. Generally speaking, any source has finite capability that manifests itself in answering easy questions with higher accuracy than difficult ones. Difficulty of various questions is source-specific: what is easy for one source can be difficult for another and vice versa.

On the other hand, information contained in an answer to any question carries a certain value of information with respect to the given decision making problem. The latter measures the improvement in the value of the problem objective resulting from the information contained in the answer. The decision maker would naturally be interested in maximizing this value of information [26] and can achieve this goal by carefully choosing a question that would be sufficiently easy for the source to yield an accurate answer and, at the same time, relevant to the problem at hand so that the resulting value of information would have the highest possible value. This naturally leads to an important question the decision maker appears to be facing: how the information source should be optimally "aligned" with the given problem, or, more precisely, what question the decision maker should ask the information source so that the respective answer would have the largest positive effect on the solution quality for the given problem. More generally, if several information sources are available the decision maker would want to know what question(s) and, possibly, in what order the sources should be asked so that the combined effect of the respective answers on the solution quality can be maximized. In other words, here the overall problem is that of optimal “alignment” of a system of information sources to the given decision making problem. What can make that latter problem more difficult is that optimal question(s) to be asked a given source might in general depend on the number and properties (“expertise”) of other available sources.

If such a methodology is to be developed, it seems logical to begin with (i) a quantitative framework describing information sources, questions and answers, (ii) study relationships between questions and the value of information of answers of the given source to these questions and (iii) use the results of (i) and (ii) to develop algorithms for choosing optimal questions and thus optimizing the process of acquiring additional information from the available source(s) for the decision making problem of interest.

1.1.3 Quantitative Aspect

The Problem

When uncertainty is present, several approaches to decision making are used depending on the problem at hand. If the main difficulty lies in a large number of possible solutions as well as a complex structure of the feasible region then optimization methods are usually used (stochastic [7], robust [6, 5]).

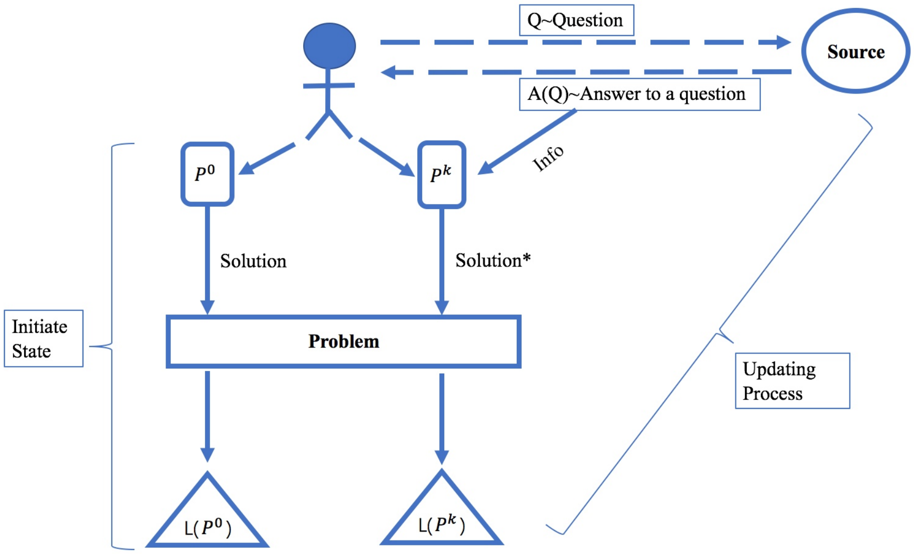
Even though the problem can be of rather general form, since the study of quantities is our goal, we will assume it to have a typical optimization under uncertainty shape. In decision making under uncertainty, the goal is to choose the best decision given the available information, according to a suitable criterion. One of the most widely used criteria is that of optimizing the expected objective function given the probability distribution that describes the available information. The problem so formulated can be formally written as

$$min_{x \in X} \ \mathbb{E}_p f(\omega, x) = \int_{\Omega} f(\omega, x) P(d\omega) \tag{1.1}$$

where Ω is the parameter space and $P(\cdot)$ is a probability measure on it. Here $X \subset D$ is the set of all feasible solutions, i.e. the set satisfying all (deterministic) constraints that are present in the problem formulation, where D is the space to which all solutions belong (e.g. a suitable Euclidean space). Ω has the meaning of a space of possible values of input data parameters that are not known with certainty. It is often referred to as a parameter space. P is a fixed initial probability measure (with a suitable sigma-algebra assumed) on Ω that describes the initial state of the uncertainty and that can in principle be modified by querying information sources. The function $f : X \times D \to \bar{\mathbb{R}}$ is assumed to be integrable on Ω for each $x \in X$. For example, in the context of stochastic optimization, X is the set of feasible first-stage solutions and $f(\omega, x)$ is the best possible objective value for the first stage decision x in case when the random outcome ω is observed.

We are interested, given the problem (1.1) and an information source capable of providing answers to our questions, in obtaining the best possible solution to problem (1.1), suitably modified by the source's answer(s). Figure blow shows this process. It is worth noting that the information exchange process between the agent and the source represented by the dotted line in the figure which belongs to the category of classical information theory. To study how to transmit this information is not within the scope of the study, we are concerned about what kind of questions should be asked to get the most efficient answer. It is necessary to assume that the information (knowledge) of the agent at the beginning is incomplete. Otherwise, the exchange of information is not required. With the general assumption as a starting point, Cox believes that the only way to

Figure 1.4: Overall Process



quantify the incomplete information and update its dynamics is through the standard probability. Thus, the initial state of the agent's information is described by the probability measure $P(\cdot)$ on Ω. The agent's goal is to update this measure using information obtained from the source. Agents need some specific information with specific content to solve the problem. In order to get it, he must describe it to the source. The description of the agent's desire to obtain information from the source will be called the **question**. Accordingly, the source's response to the question that supplies the information described in the question will be called an **answer**.

To make this desideratum a bit more specific, let $L(P)$ be the expected loss corresponding to measure P defined as follows.

$$L(P) = \int_{\Omega} f(\omega, x_P^*)P(d\omega) - \int_{\Omega} f(\omega, x_\omega^*)P(d\omega)$$

where , x_P^* is a solution of (1.1) and x_ω^* is a solution of $min_{x \in X} f(\omega, x)$for the given ω.

Let Q be the set of all possible (suitably defined) questions that can be directed towards the source of information, and let $A(Q)$ be its answer to a particular question $Q \in \mathcal{Q}$. Further, let P^k be the measure on Ω conditional on reception of a particular value a of the answer A. One can think of P^k as the measure updated by the value k, from the original measure P. Then the expected loss following question Q and answer $A = A(Q)$ can be found as

$$L(P, Q, A(Q)) = \sum_k Pr(A(Q) = k) \int_{\Omega} f(\omega, x_{P^k}^*)P^k(d\omega) - \int_{\Omega} f(\omega, x_\omega^*)P^k(d\omega) \tag{1.2}$$

where the sum is over all possible values a of the answer A.

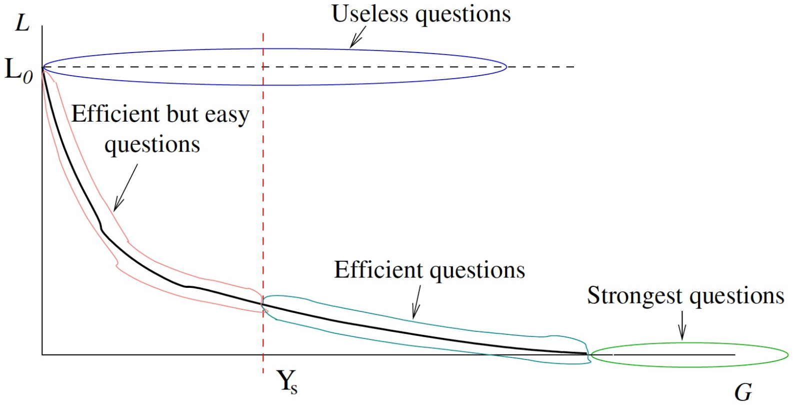
Our goal then can be stated as that of finding, for the given problem (1.1) and a given information source, the question(s) $Q \in \mathcal{Q}$ that would make the corresponding expected loss (1.2) as small as possible:

$$min_{Q \in \mathcal{Q}}\ L(P, Q, A(Q))$$

Here is a figure about how a question can be classified.

Informally speaking, the problem is about finding the question(s) that is "aligned" optimally with both the information source's "strengths" and the particular decision making problem.

Figure 1.5: Expected Loss



Changing the purely "optimization" component of the problem (the function $f(\omega, x)$ and the set X) while keeping the "information" component (the space Ω and the measure P) the same will in general change the optimal question(s) Q for the same information source. Thus the main goal can also be described as that of finding an optimal alignment between the optimization and information components of the problem (where the information source itself is included in the latter).

Information Exchange Framework

The main components of the information exchange framework developed here are information sources, decision maker's questions and corresponding source's answers. The answers change the agent's beliefs about the problem parameter space thus affecting the resulting solution quality. The answers – and the information contained in them – are generated by the source's knowledge and reflect the quality of the latter. On the other hand, the answers directly depend on the corresponding question which reflect the agent's needs – the specific detail of the agent's beliefs which are deemed either lacking in precision or important for the given problems (or both). If one wants to study the process of information exchange from a quantitative point of view one needs to make use of some suitable quantitative characteristics of the elements of the said process – questions, answers and the sources knowledge – as well as the agents original beliefs which are updated by the information received from the source.

Some preliminary results were obtained [39, 37, 38, 40]. Below, we briefly describe an updated version of these results. The main quantitative characteristics developed previously are *Question Difficulty* and *Answer Depth*. The most recent research indicates that these two characteristics have to be augmented by one more – the *Question Depth* which we also briefly introduce below.

Questions and Question Difficulty

In our current view, question difficulty is a derivative and the most complicated concept of the three, with the question and answer depth being the simpler ones. We begin with the question difficulty here partially due to "historic" reasons and partially to emphasize the "naturalness" of it from the point of view of an axiomatic characterization.

Recall that the problem was assumed to have the form

$$min_{x \in X}\ \mathbb{E}_p f(\omega, x) = \int_{\Omega} f(\omega, x) P(d\omega) \tag{1.3}$$

where Ω is the problem parameter space, and P is the probability measure on Ω encoding the initial state of agent's information.

Questions were identified in [39] with partitions $C = \{C_1, ..., C_r\}$ of the parameter space Ω of the problem. Partitions were allowed to be incomplete, i.e. such that $\cup_{j=1}^{r} C_j$. The *question difficulty* functional was introduced to measure the degree of difficulty of the question to the given information source, so that the information source would be able to answer questions with lower values of the difficulty functional more accurately that those with higher values of difficulty. The specific form of the difficulty functional was determined in [39] by demanding that it satisfy a system of reasonable postulates that, in particular, imposed the requirements of linearity and consistency. The question difficulty introduced here can be measured by the *difficulty functional*— denoted by $G(\Omega, Q, P)$. We would like to determine the overall shape of it. A standard way of doing so is to demand that $G(\Omega, Q, P)$ satisfy some requirements that express consistency and other desiderata derived from what is known about the general nature of the object. Such requirements can be formulated as postulates and the overall form of the functional can be then derived up to some remaining degrees of freedom. The latter will then express characteristics (parameters) of the particular source.

If we require the difficulty to be *additive* for questions that are successive "detalizations" of each other, the question difficulty functional has to be logarithmic in probability measure expressing the agent's initial state of information.

$$G(\Omega, Q, P) = \sum_{i=1}^{r} u_{\Omega}(Q) P(C_i) \log \frac{1}{P(C_i)} \tag{1.4}$$

if all subsets C_i belong to the same context of the source's knowledge structure. Here, $u_{\Omega}(Q)$ is the coefficient characterizing the source's knowledge depth in this particular (global - belong to the whole space Ω) context.

More generally, if the source's knowledge structure has multiple "crossed" or "nested" contexts, the question difficulty functional will reflect that.

$$G(\Omega, Q, P) = \sum_{k=1}^{p} \sum_{C \in Q^{(k-1)}} P(C) G_0(C, Q_C^{(k)}, P_C) \tag{1.5}$$

where $(Q^{(1)}, Q^{(2)}, ..., Q^{(p)})$ is some sequence of partitions of Ω such that $Q^{(k+1)}$ is a refinement of $Q^{(k)}$ for $k = 1, 2, ..., p-1$, $Q^{(0)}$ denotes the trivial partition $Q^{(0)} = \{\Omega\}$, and

$$G_0(\Xi, Q, P_{\Xi}) = \frac{\displaystyle\sum_{j=1}^{r} u_{\Xi}(Q) P_{\Xi}(C_j) \log \frac{1}{P_{\Xi}(C_j)}}{\displaystyle\sum_{j=1}^{r} P_{\Xi}(C_j)} \tag{1.6}$$

In particular, the difficulty of the given question Q depends on, besides the initial probability measure P, the function $u_{\Xi}(\cdot)$ defined on the parameter space Ω and depending on the context.

One can say that, while the source's knowledge (in its full detail) generates the agent's information change (which is practically impossible to compute), the system of contexts generates the question difficulty which is (relatively) easy to estimate. This is what allows the agents to choose between sources and make use of the provided information (since the accuracy of it is predictable). A source with deep but irregular knowledge would be impossible to acquire information from. As we have seen, in the simplest case, there is a single context for the source and the difficulty functional reduces simply to the (multiple of) Shannon entropy of the probability distribution induced by the question (partition of Ω).

Answer Depth

Here, we give a brief review of the concept of answer depth which appears to be obsolete at the time of writing. We will have to say more about this towards the end of the thesis.

Given a question $Q = \{C_1, ..., C_2\}$, the information source can provide an answer $V(Q)$, defined in [37], that takes one of values in the set $\{s_1, ..., s_m\}$. A reception of the value s_k has an effect of modifying the original probability measure P on Ω to a new (updated) measure P^k. To ensure the answer , $V(Q)$ is in fact an answer to the (complete) question Q and no more. The folloing condition is required to hold for the updated measures $P^k, k = 1, ..., m$:

$$P^k = \sum_{j=1}^{r} v_{k|j} P_{C_j} \tag{1.7}$$

where $v_{k|j}, k = 1, ..., m, j = 1, ..., r$ are the corresponding conditional probabilities.

The *answer depth* functional $Y(\Omega, Q, P, V(Q))$ for the answer $V(Q)$ to question Q measures the amount of pseudo-energy that is conveyed by $V(Q)$ in response to question Q. The general form of $Y(\Omega, Q, P, V(Q))$ can be established if certain reasonable requirements (postulates) it has to satisfy are imposed. A system of postulates proposed in [37] that parallels the postulates for question difficulty and in particular, impose the requirements of linearity and isotropy. The following theorem was then proved in [37]. The extent to which a source answers a given question can be quantified by the answer depth functional $Y(\Omega, Q, P, V(Q))$.

$$Y(\Omega, Q, P, V(Q)) = \sum_{k=1}^{p} \sum_{C \in Q^{(k-1)}} P(C) Y_0(C, Q_C^{(k)}, P_C, V(Q_C^{(k)})) \tag{1.8}$$

where $Y_0(\Xi, Q, P_\Xi, V_\Xi(Q))$ denotes the depth of answer $V_\Xi(Q)$ to question Q on subset (i.e. in the context of) Ξ on the base space. Here, the elementary depth functional $Y_0(\Xi, Q, P_\Xi, V_\Xi(Q))$ can be shown to have the following form:

$$Y_0(\Xi, Q, P_\Xi, V_\Xi(Q)) = \sum_{k=1}^{m} v_{k|\Xi} Y(\Xi, Q, P_\Xi, P_\Xi^k) \tag{1.9}$$

where P_Ξ^k is the measure modifies by the reception of $V_\Xi(Q) = s_k$ and $Y_0(\Xi, Q, P_\Xi, P_\Xi^k)$ is the conditional (single-context) depth that depends on the modified measure P_Ξ^k. The latter looks like

$$Y_0(\Xi, Q, P_\Xi, P_\Xi^k) = \frac{\displaystyle\sum_{j=1}^{r} u_\Xi(Q) P_\Xi^k(C_j) \log \frac{P_\Xi^k(C_j)}{P_\Xi(C_j)}}{\displaystyle\sum_{j=1}^{r} P_\Xi^k(C_j)} \tag{1.10}$$

Knowledge Structure

After the agent receives the knowledge of the source, his knowledge will be changed. Since the latter is quantitatively described by probabilistic measurements, the change process must follow the standard Bayesian law

$$P_a(\omega|m) = \frac{P_s(m|\omega)P_a(\omega)}{\sum_{\omega' \in \Omega} P_s(m|\omega')P_a(\omega')}, \tag{1.11}$$

where m is the particular value of the message sent to the agent by the source. We see that the knowledge of the source can be described quantitatively by the conditional distributions $\{P_s(m|\omega)\}_{\omega\in\Omega}$. If all conditional distribution $P_s(m|\omega)_{\omega\in\Omega}$ were known or a source was available for which

$$P_s(m|\omega) = \delta_{\omega,m_\omega}, \tag{1.12}$$

where m_ω is the message associated with the element ω (perfect knowledge source) then no additional theory would be needed. Similarly, if each channel is perfect, all external information forms are non-redundant, you do not need the classic information theory. In practice, however, these conditional distributions have very complex structures that involve many parameters. In most cases, it is impractical to estimate them with reasonable accuracy.

Since it is not possible to estimate the source of knowledge parameters, the agent must rely on other ways to optimize the information extraction. We can see that in practice, the agent can evaluate the various sources of information and choose between them to align the resources with their problems.The most common form of this phenomenon is to find a “expert” source, the kind of expert known to be close to the perfect knowledge in one aspect of the problem.

If you want to get information from resource optimization - like the classic information theory, manage the way in which information is optimized - the first step is to think about the source’s knowledge. Obviously, to make the source code knowledge useful to the agent, it must have some rules of the structure. Otherwise, even if the agent can get them, he will not know what to do with the source message.

The knowledge is expressed by conditional probabilities $\{P_s(m|\omega)\}_{\omega\in\Omega}$ which lives in a simplex of dimension equal to $|V|-1$ (one fewer than the number of possible values of the source's message). Let us assume that the structure of the source's message is such that the source is in principle capable of providing perfectly correct information. In other words, the source is aware of all elements of the problem parameter space Ω. Thus $|V| = |\Omega|$. Then, geometrically, the knowledge structure of the source will be represented by $|\Omega|$ points in a $(|\Omega| - 1)$ dimensional simplex. These points have to possess some regular structure in order for the source's knowledge to be usable. We call a set of such points on Ω or its projection a context if all these elements are found in the same relation to each other. In other words, the corresponding conditional distributions can all be obtained from each other by some permutation of the values of elements V.

For a real-life source, the contexts will, in general, not be exact. This is similar to regression: the exact relationship between some parameters of a system is typically less regular and more complicated but can often be described approximately by a simple function (i.e. linear) reasonably well.The contexts themselves can form a fairly complicated structure, with multiple contexts being present and "crossed", or "nested" in DOE terminology, with respect to each other.

Consider the base space Ω consisting of four elements: Green and red apples and green and red pears. The information source can be asked to distinguish between those four objects. The figures below shows the geometric view of this example. In this case, the source has two well-defined contexts in its knowledge structure: "Type" and "Color". The context "Type" is well defined because the source distinguishes green apple from green pear just as well as he does red apple from red pear. The question about the color will be more difficult for the source than that about type. The value of the respective knowledge structure parameters will be respectively higher. Figures about Type and Color are shown below.

Now let's consider a slightly different knowledge structure. The source is still an expert in fruit type but can distinguish the color of pears better than that of apples. Thus Color is not a global context for this source. Instead he has "nested" color contexts: different for each type.

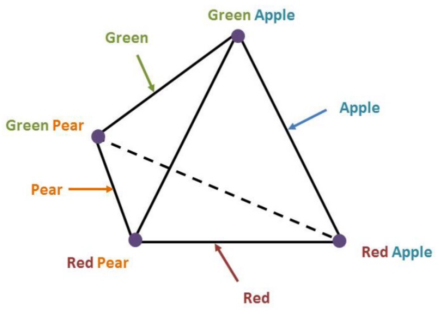


Figure 1.6: Example: Apple and Pear

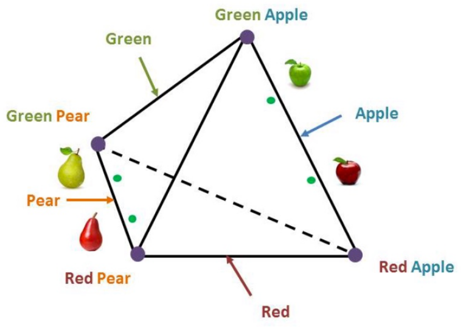


Figure 1.7: Elements of Apple and Pear

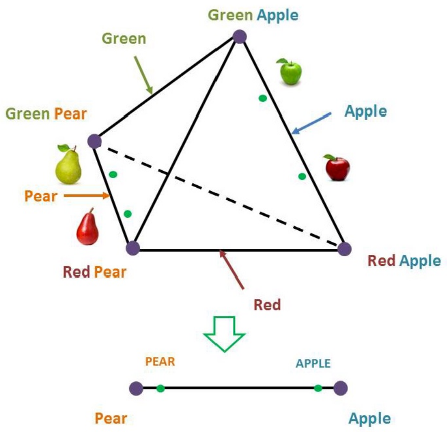


Figure 1.8: Type Context

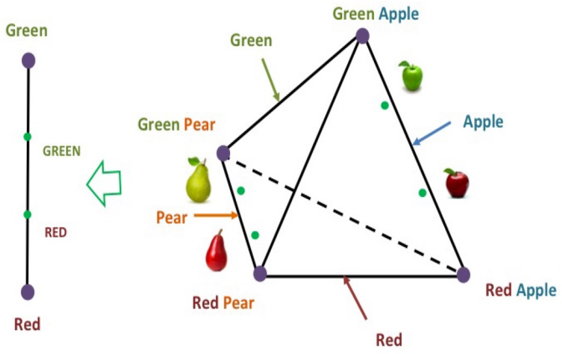


Figure 1.9: Color context

Now recall the function of question difficulty $G(\Omega, Q, P)$ and answer depth $Y(\Omega, Q, P, V(Q))$:

$$G_0(\Xi, Q, P_\Xi) = \frac{\sum_{j=1}^{r} u_\Xi(Q) P_\Xi(C_j) log \frac{1}{P_\Xi(C_j)})}{\sum_{j=1}^{r} P_\Xi(C_j)}$$

and

$$Y_0(\Xi, Q, P_\Xi, P_\Xi^k) = \frac{\sum_{j=1}^{r} u_\Xi(Q) P_\Xi^k(C_j) log \frac{P_\Xi^k(C_j)}{P_\Xi(C_j)}}{\sum_{j=1}^{r} P_\Xi^k(C_j)}.$$

The coefficients $u_\Xi(\cdot)$ that enter the difficulty and depth functionals are related to the source's knowledge structure.

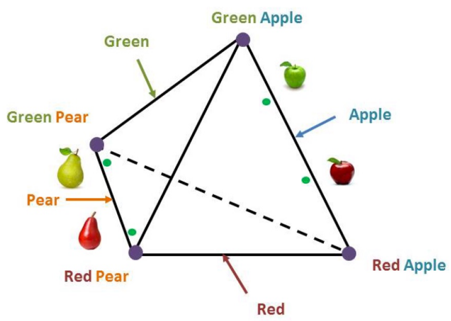


Figure 1.10: Example 2: Type Context

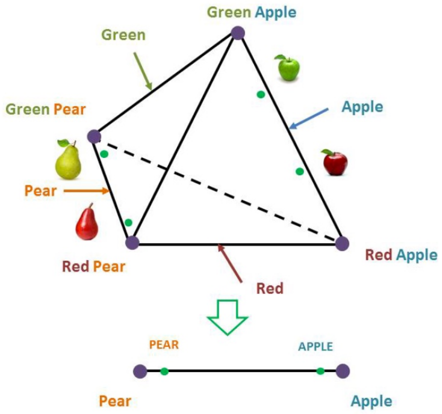


Figure 1.11: Example 2: Color Context

Revised answer depth and question depth

The answer depth functional described above belongs to the preliminary version of the proposed approach. It was later found that such definition of answer depth leads to somewhat inadequate description of the information exchange process. Here we briefly sketch the revised version of the answer depth together with the additional function – the question depth.

First, let's turn our attention to ways of relating the coefficients $u_\Xi(\cdot)$ to the source's knowledge structure. Since the question difficulty functional is linear in $u_\Xi(\cdot)$, a natural way to express this relation is to set $u_\Xi(\cdot)$ higher for the case of "weaker" source thus making the corresponding question would more difficult to the source. To see how it works, let's consider a single context Ξ. Let $v_{k|j} = Pr(s = s_k|\omega_i)$ denote the probability of the message having value s_k when the true state of the system is described by ω_i. Let $v_{k|\Xi}$ denote the unconditional (within the subset Ξ) probability of the sources's message taking the value s_k. Then the depth of the answer to an ideal (single subset) question ω_i such that $P(\omega_i) = \frac{1}{r_\Xi}$ (thus describing the "maximum ignorance" of the agent) will be equal to $u_\Xi(\omega_i) \sum_k v_{k|i} log \frac{v_{k|i}}{v_{k|\Xi}}$, where the value of $u_\Xi(\omega_i)$ will be the same for all ω_i within the same context inside Ξ.

We can set the value of $u_\Xi(\cdot)$ by using the convention that the maximum answer depth the source is capable of within the context Ξ is equal to $logr_\Xi$ where r_Ξ is the number of distinguishable elements n Ξ. This can be achieved by setting

$$(u_\Xi(\omega_i)+1)\sum_k v_{k|i} log \frac{v_{k|i}}{v_{k|\Xi}} = log r_\Xi$$

for all $\omega_i \in \Xi$. Clearly, if the source knowledge on Ξ is perfect one would obtain $u_\Xi(\omega_i) = 0$. Otherwise, $u_\Xi(\omega_i) > 0$.

A question difficulty thus defined will always vanish for a source possessing perfect knowledge of what's being asked. Moreover, such question difficulty (including the more general multiple context case) will have the meaning of the average value of $u_\Xi(\cdot)$ over all contexts involved weighted with the degree of "ignorance" of the agent about these contexts. This also gives us a hint of how one should naturally define the answer depth. Specifically, the revised version of the answer depth functional doesn't use the quantities $u_\Xi(\cdot)$ expressing the source's knowledge structure and simply measures the extent of the change of the agent's beliefs upon the answer reception (more on this later).

This naturally leads to the notion of question depth defined simply as the largest value the answer depth can take in case the corresponding answer is perfect. It's straightforward to show that the question depth thus defined will have the form of simply

$$D(\Omega, Q, P) = \sum_{i=1}^{r} P(C_i)\log\frac{1}{P(C_i)},$$

which is nothing else but the Shannon entropy of the probability measure induced by the partition representing question Q. The question depth thus defined is a measure of the initial ignorance of the agent with respect to the question the agent is asking. The question depth is higher if either the question is very detailed or the agents knows little about the matter. A high value of question depth gives the source the opportunity to change the agent's beliefs by a large amount – measured by the revised answer depth functional. On the other hand, if the question depth is low the agent's updated belief will stay close to the original – no matter how much the source knows.

1.2 Motivation and Background

1.2.1 Motivation

The main motivation for this thesis is the need to revisit the existing quantitative measures of information in order to clarify certain points of the overall proposed approach and resolve some of the contradictions of the preliminary version of it which were recently found.

To this effect, the logical place to begin is to clarify what information really is – at least in a more narrow sense directly related to optimizing the process of its acquisition in science and engineering. We propose to use the following definition.

Definition 1.2.1. *Information is an (ideal) image of some human activity.*

In this definition the word “ideal” refers to some material entity which is used to represent another material entity, like, for example, symbols that are used to record and transmit information are there solely to represent something else – the information content.

The need by the agent to acquire additional information is only there if the information he or she has at the time is incomplete. Thus one would need to quantify incomplete information first. Fortunately Cox had argued [17, 18] that the only consistent way to quantify that is by describing it by standard probabilities subject to Bayes’ update rules (together with possible generalizations (see [12] for details)).

In a nutshell, Cox’s argument amounts to the observation what complete information can be stated as a system of logical propositions which can be either true of false. For example if the agent has complete information about the apple color (knowing for sure that it’s red) than the proposition “The apple is red” is true and the proposition “The apple is green” is false. On the other hand, if the agent’s information is incomplete then neither of these propositions can be true or false. This implies that either of them has to be partially true. Since they are mutually exclusive if one of them is almost true the other has to be nearly false and vice versa. This leads to the notion of quantitative difference between qualitatively uniform (partially true) propositions. The corresponding quantity has to be additive with for mutually exclusive propositions. Then, it is easy to show that consistency with logical algebra requires that this quantity be multiplicative with respect for logical intersections. This uniquely leads to an identification of these quantities (degrees of belief) and classical probabilities.

Once one is able to describe – qualitatively and quantitatively – the state of agent's information (belief) as incomplete, the next task is to find suitable quantification of the degree of change of these beliefs. Clearly, the original beliefs might change just a little bit or very drastically. The important question is what is the correct quantity – if it exists – characterizing such a change. Since – according to some authors – information is "whatever can change rational beliefs" [12] (the property of information that follows directly from our definition), the degree of such change can be identified with the true quantity of information obtained. It is this quantity which constitutes the main subject of the present thesis.

1.2.2 Background

The field of Information Theory, born from Shannon's work on the theory of communications [46] has had great success in a number of fields — besides communications itself which it revolutionized — that include statistical physics [28, 29], computer vision [50], climatology [36, 49], physiology [31] and neurophysiology [13]. The relatively new field of Generalized Information Theory (see e.g. [32]) is concerned with problems of characterizing uncertainty in frameworks that are more general than classical probability such as Dempster-Shafer theory [45]. There it was shown, for example, [35, 24] that the minimal uncertainty measure satisfying consistency requirements (such as general subadditivity and additivity for combining uncertainty for independent subsystems) is obtained by maximizing Shannon entropy over all classical probability distributions consistent with the given (generalized) belief specification.

In our research, together question difficulty and answer depth can be thought of as a logical development of the entropy concept of information theory. The axiomatic approach was first used, besides Shannon himself, in [20] to derive the most general form of the entropy function. Later, [43] used a different set of axioms to find the one-parameter family of functions (later called Rényi entropies) that included standard (Shannon) entropy as a special case. The concept of structural entropy was introduced in [25] and used for classification purposes. Also known as Havrda-Charvat entropy, it was more recently obtained by axiomatic means in [47] where axiomatization of partition entropy was discussed on rather general grounds (see also [27] for closely related work). It was shown in [47] that Shannon entropy, Havrda-Charvat entropy and Gini index all obtain as particular cases of general partition entropy that satisfies a system of reasonable axioms.

Similarly, it can be thought of as a development of a general theory of inquiry that goes back to the work of Cox [17, 18]. This line of work received more attention recently resulting in a formulation of the calculus of inquiry [33] that constructs a distributive lattice of questions dual to the Boolean lattice of logical assertions. The definition of questions adapted in Chapter 2 corresponds to the particular subclass of questions — the partition questions ? defined in [33]. The work here goes beyond that on the calculus of inquiry in that it introduces the concept of pseudo-energy as a measure of source specific difficulty of various questions to the given information source. One could say that it develops a quantitative theory of knowledge as opposed to the theory of information.

Explicit consideration of information sources that lies at the core of the proposed methodology is similar in spirit to analyzing and using information provided by human experts. In fact, in many practically relevant applications the role of multi-purpose information sources used in the proposed approach will likely be played by experts. In existing research literature, the problem of optimal usage of information obtained from human experts has been addressed mostly in the form of updating the decision makers' beliefs given probability assessment from multiple experts [22, 23, 14, 15] and, in particular, optimal combining of expert opinions, including experts with incoherent and missing outputs [42]. Closely related to the approach initiated here are the investigations on using and combining information of experts that partition the event differently [8] and on rules of updating probabilities based on outcomes of partially similar events [9]. The latter investigations essentially consider experts that provide qualitatively different information. The dependence of the quality of experts' output on the particular partition was also studied in [21]. Here, the emphasis is on optimizing on the particular type of information (i.e. partition) for the given expert(s) and the given decision making problem.

1.3 Information Measures

In order to understand the idea of information, we began to consider some ideas about the concept of information; our daily experience tells us that we continue to deal with information and information transmission. One reason for this information is that all creatures protect their lives by receiving information about their environment so that they can discover food and danger ("Information is the very essence of life; H.Haken"). Despite the importance of information, we have not yet found the definition of this concept. One of the reasons for this lack of definition is that when we check information about messages, objects, etc., we need different levels. For example, the mathematical description of the information used in communication theory can only provide a limited number of levels. Thus, as described by mathematical measures, the concept of information can provide us with an in-depth understanding of the whole concept of information. However, these limitations are due to the neglect of the subjective information, which is proven by common scientific methods, trying to eliminate human beings from their experiments as an immeasurable source of error. Such subjective aspects of information may be complex, or very inaccurate models.

A possible way out could be the use of the effect caused by the received information and the description of the information processing using this effect. This can be done easily by sending a message to someone and then by judging his reaction. However, this analog information transmission already contains some different information processing steps, merely describing the merging of these different information processing steps. After that, we need to be evaluated by certain criteria, and we must judge the comparison between intent and response. However, this is an easy task because of the objective criteria that can be applied. But the key to information transmission is the steps that must be inserted and the conversion of the information that must be completed in this transmission. It is in these steps that we need to describe the details, which are the basis for the actual implementation of modeling. Have the opportunity to optimize individual details in the right way, and the entire system will benefit from this optimization. However, this requires understanding the concept of information, allowing at least a limited definition of information. When we can define at least part of the concept of information, we can also use this part of the understanding to achieve the application, so as to learn from the information collected on the information about more information. So we can extend our understanding of this abstract concept by constantly examining our actions in information. If we only optimize the system for the transmission rate, we do not need to consider the concept of information, because the boundary conditions (bandwidth, time and energy) impose restrictions on the expected optimization, this optimization must be adjusted conditions. There is no need to understand the concept of information to do such an optimization. So the advantage of this aspect is that we can build and optimize the information transmission system, although we do not have a complete understanding of the actual information.

The complexity of the information is clearly a great challenge to the definition, which limits the scope of application of this concept. However, we can understand some parts of conceptual information and apply it to scientific theory (coding theory, estimation theory, cryptography, ...). Thus, it is possible to develop a scientific information theory, which assumes the comprehension of particular facts and allows us to use this comprehension.

We, therefore, deal with the concept of information in an emotional way by benefiting from the information available in our limited field of information from the available and reproducible experience. Because this inductive strategy is based only on partitions that can be experimentally or experimentally verified, the resulting theory must be continually developed to include all the changes that are collected through additional experience. Within the scope of these experiences, over the past 60 years, many other options have been developed for information and communication system information description. They all expand the existing theory, because they are more numerous, and even provide a complex accumulation of completely different functions. It is, therefore, difficult to determine whether the specified function describes, understands, or even loves a given question in the desired way. However, this diversity is the result of the induction method and can not lead to a complete description of the problem due to the necessary limitations of the particular case.

By fully adapting the boundary conditions, the deduction from the description of the complete topic to the description of some limited areas will require the characterization of the information, which is not suitable for mathematical selection. But this is not necessarily a negative attribute of information because every math system is incomplete, so we can not use our resources to prove all the statements of a mathematical system. If you change the description level of the system, you can get other insight that has not been achieved before. Therefore, different levels of information complement each other; we use these opportunities to verify and explain our information in the mathematical processing of the results obtained.

1.4 Objective of Thesis

The main objective of this thesis is to study, review and find a suitable methods of information measurements to support our current research introduced in Section 1.1. Specifically, this thesis tries to find a proper aspects of "Information" that meets our definition of information and discover right information measures among existing methodologies that can be used in the quantitative aspects of our research.

1.5 Thesis Outline

Chapter 1 is the introduction of this thesis. In section 1.1, the general ideas in the research of our group is introduced with some detailed information like overfall framework and quantitative aspects that can be useful as the standard to distinguish different information measures in following chapters. Section 1.2 shows the motivation and background of this thesis. It includes several questions we want to answer about information and some previous work that has been done. Section 1.3 introduces the core concept of this thesis about information measures, about the way we treat "information". Section 1.4 introduces the objective of this thesis and Section 1.5 shows the structure of it.

In chapter 2, I will examine some historical milestones in the development of information theory and arrive at Shannon's Information, which is the fundamental one. The character of this measure of information is more than that of an entropy which has been known in physics.

In chapter 3, I will introduce another famous measure of information—Rényi's Information. In particular, Shannon's information is a special case of Rényi's Information with $\alpha = 1$. We will focus on the generalized mean of entropy and see how different measures make different characterizations with different properties.

In Chapter 4, there is a critical measurement of information derived — Kullback-Leibler Divergence. It can be called in different names like Kullback's information and Relative Information. The key point is Kullback-Leibler Divergence is a good measure to deal the difference between two information/entropies.

In Chapter 5, a summary of popular information measures is provided. And the relationships among Information, Entropy and Probability are discussed. Based on the quantitative aspects in our research, we will provide a brief introduction about the deduction of our main idea that what kinds of information measures are necessary.

In Chapter 6, we will see different divergence functions in the view of information geometry. By introducing some important properties of Bregman Divergence and Invariant Divergence, we show how the requirements of our ideal information measure are satisfied by these two classes of divergences. And since Kullback-Leilber Divergence is the only one that belongs to both Bregman divergences and Invariant divergences, it is indeed the proper information measure we are seeking for our research.

Chapter 7 is the summarization of this thesis. Conclusions achieved in this thesis are derived in section 7.1. And a brief discussion about the future work is involved in section 7.2 with some potential difficulties.

Chapter 2

Information Theory and Shannon's Information

2.1 Information Theory

2.1.1 Information Transmission

Samuel F.B. Morse 1837

The Morse alphabet was developed to transmit messages in an optimal way over a communication channel. The assignment of the characters of the alphabet to the symbols, which was carried out by S. Morse, was not based on the relative frequency with which the letters appeared in given texts. It was instead carried out by S. Morse by counting the types in a letter case of the typesetter in a print office. Thus the Morse alphabet is not the optimum coding on the basis of the preliminaries and it can be improved by a factor of 15% [41].

With this coding the rate of transmission of a message is bounded by the divergence, leading to a merging of two subsequent pluses on the channel. A line requires three times the periods of a dot and thus prevents the fast transmission of messages, which was the reason to replace the line by a dot with a negative sign. This leads to an increased speed of transmission.

Thomas Edison 1874

Thomas Edison introduced the quadruplex coding, using the symbols $+3, +1, -1, -3$ to increase further the rate of information transmission. The amount of information that can be transmitted over a given channel obviously depends on the following three factors:

- How fast can we send two successive characters without leading to a merging of both on the channel(divergence, inter-symbol interference)? —frequency, bandwidth
- How many different characters are we able to transmit (number of levels)? — amplitudes
- How much time do we have for the transmission? — time

Nyquist 1924

in his paper 'Certain factors affecting telegraph speed' in [41], Nyquist says:

> If we send symbols at a constant rate, the rate of transmission W is connected to the number of possible different symbols that can be sent, by the equation
>
> $$W = B \cdot \ln m \tag{2.1}$$
>
> Where B is a constant, depending on the number of different amplitudes of current that can be sent per second; m is the number of different symbols at our disposal for the transmission.

2.1.2 Information Functions

Hartley 1928

Hartly defined a unit and a measure of information by stating the following prediction:

> The answer to a question that can assume the two values 'yes' or 'no' (without taking into account the meaning of the question) contains one unit of information.

The unit of information is called a bit, because we can realize the two answer in a dual system 0 and 1. Now we are also able to realize complex systems or procedures by an arbitrary number of these binary digits. The characterization of a set E_N consisting of $N = 2^n$ dual enumerated elements, where the occurrence of the elements can be answered with yea or no, requires the information

$$I(E_N) = n = 1\log N = \frac{\ln N}{\ln 2} \text{ bit} \tag{2.2}$$

This definition of information is commonly known as Hartley's formula and it holds under the assumption that all N elements of the set E_N occur with equal probability ($p_{n_i} = \frac{1}{n}, for\ i = 1, ..., N$). As Hartley' formula is a definition, we need not perform any proof. But this definition is not at all an arbitrary selection. It meets the following postulates of Hartley summarizing the requirements of an information function.

For the sets $E_N, E_{N+1}, E_M, E_{MN}$, consisting of $N, N+1, M, MN$ equal probable elements we postulate:

- $I(E_{MN}) = I(E_N) + I(E_M)$

  The sum of the pieces of information of two independent sets E_N and E_M is equal to the information of the union set E_{MN} (all sets consist of elements occurring with equal probability).

- $I(E_N) \leq I(E_{N+1})$

  Information is a monotonically increasing function.

- $I(E_2) = 1$

  Determination of a unit quantity to obtain an absolute reference.

Hartley's information (information content)

$$I(E_N) = n = 1dN = \frac{\ln N}{\ln 2} \text{ bit} \tag{2.3}$$

and the channel capacity

$$C = \log n \tag{2.4}$$

are identical, though they are defined with different preliminaries. Hartley examined the probability of the occurring events, while the channel capacity aims at the number of events to define a quantum to measure the quality of a signal processing system or a communication system. Since for equally probable events the reciprocal probability is equal to the number of occurring events, both definitions are identical an we thus get:

$$Hartley's\ information\ =\ channel\ capacity$$

The next scientists to deal with the problem of information theory were Norbert Wiener and Andrej N. Kolmogorov. They studied the problem of predicting a process from contaminated data and developed independent solutions for this problem. Furthermore, the evolution of probability theory by Ricard von Mises and Anderj N. Kolmogorov was an essential prerequisite for the development of the following information measures, which are merely defined on the basis of relative frequency of the probability of the occurring events.

2.2 Shannon's Information

2.2.1 Shannon 1948

Claude E. Shannon enlarged Hartley's concept of information, by permitting sets which need not occur with equal probability. The starting point of Shannon's derivation is the mutually disjoint sets $E_1, E_2, ..., E_n$ with $E = E_1 + E_2 + E_3 + ... + E_n$. The number of elements contained in the set E_k is N_k, which means that the union set E consists of $N = \sum_{k=1}^{n} N_k$ elements. As a consequence of these preliminaries, the set E_k occurs with the relative frequency

$$p_k = \frac{N_k}{N} \tag{2.5}$$

The elements of the set E_k all have the same probability of occurrence. Thus, if we know that a certain element is a member of the set E_k, the exact characterization of an element of the set E merely requires the average information I_{binary}, when we already know that the element is a member of the set E_k

$$I_{binary} = \sum_{k=1}^{n} \frac{N_k}{N} \cdot 1 \log N_k = \sum_{k=1}^{n} p_k \cdot 1 \log(N \cdot p_k) = 1 \log N + \sum_{k=1}^{n} p_k \cdot 1 \log p_k \tag{2.6}$$

thus it follows that

$$1\log N = I_{binary} - \sum_{k=1}^{n} p_k \cdot 1\log p_k \tag{2.7}$$

The information required to characterize an element of the set E is an amalgamation of two pieces of information

$$1\log N = I_{binary} + I_{Shannon} \tag{2.8}$$

$I_{Shannon}$ specifies the set E_k that contains the element; I_{binary} specifies the particular element in the known set E_k.

Shannon's entropy(information entropy) is thus given by:

$$I_{Shannon} = \sum_{k=1}^{n} p_k \cdot 1\log(\frac{1}{p_k}) = -\sum_{k=1}^{n} p_k \cdot 1\log p_k \tag{2.9}$$

Shannon's information is an expectation value of Hartley's formula $E\{1\log(\frac{1}{p_k})\}$, where the probabilities are not required to be equal for all possible events. If the probabilities are equal for all events $p_k = \frac{1}{n}$, Shannon's information reduces to the formulation of the Hartley information, which is

$$I_{Shannon} = \sum_{k=1}^{n} \frac{1}{n} 1\log n = \frac{n}{n} \cdot 1\log n = 1\log n \tag{2.10}$$

Taking a further look at Shannon's entropy, commonly denoted by $H(X)$, i.e. the entropy of the random variable X, we are usually going to use the notation H or the previously used notation $I_{Shannon}$

$$H = H(X) = -\sum_{k=1}^{n} p_k \cdot \ln p_k \tag{2.11}$$

2.2.2 Properties of Shannon's Information, Entropy

Shannon's information

$$I_{Shannon} = \sum_{k=1}^{n} p_k \cdot \ln(\frac{1}{p_k}) = -\sum_{k=1}^{n} p_k \cdot \ln p_k \tag{2.12}$$

achieves its maximum when all n symbols occur with the same probability $p = p_k$. To determine the maximum, we additionally have to take into consideration that the sum of all probabilities is equal to 1, then we have

$$\sum_{k=1}^{n} p_k = 1 \tag{2.13}$$

and

$$p_k = \frac{1}{n} \tag{2.14}$$

and this value has to be inserted into Shannon's information to determine the maximum. It is thus equal to

$$max\left(I_{Shannon}\right) = -\sum_{k=1}^{n} \frac{1}{n} \cdot \ln\left(\frac{1}{n}\right) = n \cdot \frac{1}{n} \cdot \ln\left(\frac{1}{n}\right) = \ln\left(\frac{1}{n}\right) \tag{2.15}$$

The maximum of Shannon's information is thus achieved when all events occur with the same probability $\frac{1}{n}$. But this is exactly the formulation that Hartley found to express the information. Therefore Hartley's information maximize Shannon's information.

Shannon's information describes exactly the uncertainty eliminated by the measurement. Without any observation we can give only some vague description of the possible realizations by assigning probabilities to the possible events. Having obtained an observation, the a priori uncertainty vanishes, as we are able to determine one event as the result of our measurement. Thus Shannon's information measures the information contained in an observation or the eliminated uncertainty.

The original notation of this quantity was entropy. Shannon's measure describes the initial uncertainty that we have before we make the observation. We then only know the distribution density of the experiment, as we do not have a concrete realization. But he distribution density enables us to calculate the uncertainty(the entropy) that we have regarding the desired realization. So Shannon's measure is a measure of entropy, when we look at the experiment on the basis of the initial knowledge.

Shannon's measures is also a measure of information, because by the description of the initial uncertainty it also describes the information that we are able to obtain by the following observation. When we examine Shannon's measure from this point of view, which is more with the tendency to a future gain of information, we may also speak of Shannon's information, as it describes the information contained in the succeeding observations.

Here we have to distinguish between discrete probabilities and continuous distribution density functions, because the discrete probabilities are assigned to countable events, which my or may not occur. It is thus possible to find an exact determination of the occurrence of an event by a simple yes/no decision, elimination all uncertainty regarding the realization of the observed discrete random variable. The description of a random experiment with continuous distribution densities on the other hand does not lead to an exact realization, because we are not able to perform any exact measurement. We still retain an uncertainty after we make the observation. In this case the gain of information that we get from the measurement not only depends on the initial uncertainty (as it is in the case for discrete random variable), but also on the uncertainty remaining after the measurement(which is equal to zero in discrete case).

Algebraic Properties of the Shannon Entropy

Theorem 2.2.1. *The Shannon entropies $H_n(X)$ defined by (2.11) have the following properties[1]:*

- *Symmetry*

$$H_n\left(p_1, p_2, ..., p_n\right) = H_n(P_{k(1),k(2),...,k(n)})$$

  for all $(p_1, p_2, ..., p_n) \in \Delta_n$, where k is an arbitrary permutation on 1,2,...,n.

- *Normality*

$$H_2\left(\frac{1}{2}, \frac{1}{2}\right) = 1$$

- *Expansibility*

$$\begin{aligned} H_n(p_1, p2, ..., p_n) &= H_{n+1}(0, p_1, p_2, ..., p_n) = ... \\ &= H_{n+1}(p_1, p_2, ..., p_k, 0, p_{k+1}, ..., p_n) = ... \\ &= H_{n+1}(p_1, p_2, ..., p_n, 0) \qquad (k = 1, 2, ..., n-1) \end{aligned}$$

- *Decisivity*

$$H_2(1,0) = H_2(0,1) = 0$$

- *Strong additivity*

$$\begin{aligned} &H_{mn}(p_1q_{11}, p_1q_{12}, ..., p_1q_{1n}, p_2q_{21}, p_2q_{22}, ..., p_2q_{2n}, ..., ..., p_mq_{m1}, p_mq_{m2}, ..., p_mq_{mn}) \\ &= H_m(p_1, p_2, ..., p_m) + \sum_{j=1}^{m} p_j H_n(q_{j1}, q_{j2}, ..., q_{jn}) \\ &for\ all (p_1, p_2, ..., p_m) \in \Gamma_m\ and\ (q_{j1}, q_{j2}, ..., q_{jn}) \in \Gamma_n, j = 1, 2, ...m \end{aligned}$$

- *Additivity*

$$\begin{aligned} &H_{mn}(p_1q_1, p_1q_2, ..., p_1q_n, p_2q_1, p_2q_2, ..., p_2q_n, ..., ..., p_mq_1, p_mq_2, ..., p_mq_n) \\ &= H_m(p_1, p_2, ..., p_m) + H_n(q_1, q_2, ..., q_n) \\ &for\ all (p_1, p_2, ..., p_m) \in \Delta_m\ and\ (q_1, q_2, ..., q_n) \in \Delta_n \end{aligned}$$

- *Recursivity*

$$\begin{aligned} H_n(p_1, p_2, p_3, ..., p_n) &= H_{n-1}(p_1 + p_2, p_3, ..., p_n) \\ &+ (p_1 + p_2) H_2\left(\frac{p_1}{p_1 + p_2}, \frac{p_2}{p_1 + p_2}\right) \\ for\ all\ (p_1, p_2, p_3, ..., p_n) &\in \Gamma_n,\ for\ which\ p_1 + p_2 > 0. \end{aligned}$$

Proof is omitted here. The **Symmetry** property means intuitively that in the order of events, the amount of information is constant. The **Normality** tells that experiments with two outcomes with equal probability will provide a unit of information. The **Expansibility** makes sure that the additional results with probability of 0 will not change the uncertainty of the experimental results . The **Devisivity** says in an experiment if there are two results, the probability of one is 1 and the other probability is 0, then there is no uncertainty in this experiment. The **Additivity** property describes that the information expected from two independent experiments is the sum of the informations expected from the individual experiments. And The **Strong additivity** describes the situation in which the two experience are not independent. The **Recursivity** is important. It says that by dividing an event of the system into two conditional probability events, the uncertainty of this division increases the uncertainty of the system, but only when the original event occurs. The importance of recursivity is that it gives a recursive formula for calculating H_n.

2.2.3 Application of Shannon's Information

Now let's take a look at a typical application of Shannon's information. Considering the objective of this thesis is to summarize not only the definition of each information measures but also the application fields, it is important to provide an typical example here to describe the application.

In coding theory, Shannon's information is used to construct construct codes with optimal lengths of the codewords. One of the basic equations in coding theory is the Kraft inequality, which will now derive.

If we are able to arrange message or codewords in a code tree, then we are able to state something about the number of codewords used to code the given message. Let:

1. D= number of symbols in the coding alphabet

2. $n_1, n_2, ..., n_m$ = given set of positive integers

Kraft's Inequality

$$\sum_{k=1}^{m} D^{-n_k} \leq 1 \tag{2.16}$$

is a necessary and sufficient condition for the existence of M codewords, corresponding to the end points of a tree. Their length is equal to the given numbers n_k (= length of the codewords). Proof of Kraft's is omitted here since it is pretty well-known and is out of scope of this thesis. For detail please refer to lectures or books about information theory such as [16]. It is important that the summation is done only over all disjoint codewords. There must not be any part of a codeword in the summation or a multiple summation of short codewords.Kraft's inequality is thus a sufficient condition for the existence of a set of codewords, where the length of these codewords are specified in another set. It is possible to generate prefix codes with a minimal average length, when the Kraft inequality holds.

The average of the optimum length of the codewords can be derived by

$$L = \sum_i p_i \cdot l_i = -\sum_i p_i \frac{\ln p_i}{\ln D} = -\sum_i p_i \cdot \log_D p_i = H_D(X)$$

$$L = H_D(X) = -\frac{1}{\ln D} \sum_i p_i \cdot \ln p_i = \frac{H(X)}{\ln D} \quad (\ln D^L = H(X)) \tag{2.17}$$

The average of the length of the codewords is thus equal to the entropy divided by the natural logarithm of the number of symbols that our alphabet consist of.

Chapter 3

Rényi's Measures of Information

3.1 Rényi's Information 1960

The starting point of the derivation [44] of Rényi's α-information (α-entropy) S_α and the G-divergence G_α is Hartley's definition of the information

$$I_{Hartley} = -\ln p \tag{3.1}$$

with known properties, such as additivity of the information for independent events. This definition only holds when we have no a priori knowledge regarding the probabilities p_k of the single events A_k, i.e. when all $p_k = p$ have equal probabilities. Shannon extended this description of information by assigning different probabilities to the events and calculating the average of all probabilities occurring in the observed process.

$$I_{Shannon} = \sum_{k=1}^{n} p_k \cdot \ln\left(\frac{1}{p_k}\right) = -\sum_{k=1}^{n} p_k \cdot \ln p_k \tag{3.2}$$

p_k = weighting factors of the information $I_k = \ln p_k$. The linear mean value, however, is not the only possible mean value. So Alfred Rényi extended the averaging weights $p_1, p_2, ..., p_n$ (with

$0 \leq p_k \geq 1$ and $\sum_{k-1}^{n} p_k = 1$) to the real numbers $x_1, x_2, ..., x_n$ with results in a generalized mean.

$$m_g = \varphi^{-1} \left[\sum_{k=1}^{n} p_k \cdot \varphi(x_k) \right] \tag{3.3}$$

φ is an arbitrary continuous and strictly increasing or decreasing function defined on the set of the real numbers. These generalized means are also known as Kolmogorov-Nagumo functions of the mean. If we use the generalized mean instead of the linear mean in the definition of Shannon's Information, we obtain the equation

$$I_{\text{Rényi}} = \varphi^{-1} \left[\sum_{k=1}^{n} p_k \cdot \varphi \left(1 \log \frac{1}{p_k} \right) \right] \tag{3.4}$$

or with the natural logarithm

$$I_{\text{Rényi}} = \varphi^{-1} \left[\sum_{k=1}^{n} p_k \cdot \varphi \left(\ln \frac{1}{p_k} \right) \right] \tag{3.5}$$

It is better to limit the description to the natural logarithm, as all other logarithms only introduce an additional constant factor. If $\varphi(x)$ is a linear function, the generalized mean reduces to the linear mean and we obtain Shannon's information as as a special case of this formula.

We may now use $I_{\text{Rényi}}$ as an information measure, but we certainly cannot use an arbitrary function $\varphi(x)$ with the restriction already stated. The function $\varphi(x)$ has to be selected in such a way that certain postulates are met by the resulting information function. The most important postulate is the requirement that the information of independent events can be added, to obtain the information of the union set of both events. Here Rényi uses the function

$$\varphi(x) = 2^{(1-\alpha)x} \quad \alpha \neq 1 \text{ if we use the logarithm with the base 2} \tag{3.6}$$

If we consider the natural logarithm with the base e

$$\varphi(x) = e^{(1-\alpha)x} \quad \alpha \neq 1 \tag{3.7}$$

Now we insert the generalized mean of Rényi and after several calculating, we can obtain

$$I_{\text{Rényi}} = \frac{1}{1-\alpha} \cdot \ln \left[\sum_{k=1}^{n} p_k^{\alpha} \right] \tag{3.8}$$

And if we use the binary logarithm instead of the natural logarithm, we get:

$$I_{\text{Rényi}} = \frac{1}{1-\alpha} \cdot \frac{\ln \left[\sum_{k=1}^{n} p_k^{\alpha} \right]}{\ln 2} \tag{3.9}$$

and the change in the basis of the logarithm leads only to an additional constant factor in the information function. We are thus going to use the natural logarithm, which does not require the permanent writing of factors, which do not change the meaning of the function.

3.2 Properties of Rényi's Entropy

1. Monotonicity

At first Rényi's entropy is a monotonically decreasing function of the additional parameter α. Proofs are omitted here and can be referred in [1].

2. Limits in the interval $0 \leq \alpha < \infty$

As we saw in the previous property Rényi 's α-entropy is a monotonically decreasing function of the parameter α. Thus we now want to determine the limits, which can be achieved by variation of the parameter in the interval $[0, \infty)$.

For a given distribution density $P = (p_1, p_2, ..., p_n)$ the term

$$S_\alpha = \frac{1}{1-\alpha} \cdot \ln \left[\sum_{k=1}^{n} p_k^{\alpha} \right] \quad \textit{with} \;\; \alpha \neq 1 \;\; \textit{and} \;\; \alpha > 0 \tag{3.10}$$

is a monotonically decreasing function of the parameter α with the limits

$$\ln n \geq \frac{1}{(1-\alpha)} \cdot \ln \left[\sum_{k=1}^{n} p_k^{\alpha} \right] \geq -\ln p_{max} \tag{3.11}$$

3. Nonnegative for discrete events

For discrete events Rényi's entropy is always nonnegative, which can be verified via the following consideration:

$$I_{\text{Rényi}} = \frac{1}{1-\alpha} \cdot \ln \left[\sum_{k=1}^{n} p_k^{\alpha} \right]$$
$$I_{\text{Rényi}} = \frac{1}{1-\alpha} \cdot \ln \left[\sum_{k=1}^{n} p_k \cdot p_k^{\alpha-1} \right]$$
$$I_{\text{Rényi}} = \frac{1}{1-\alpha} \cdot \ln \left\{ \sum_{k=1}^{n} p_k \cdot \exp[-(1-\alpha) \cdot \ln p_k] \right\}$$

where α is always positive. We exclude the special case $\alpha = 1$, because in this case we obtain Shannon's information, which is a special case of Rényi's information(for the linear mean).

It does not in matter in which interval the parameter α lies, the resulting Rényi information is always nonnegative. For the limit $\alpha \to 1$ we obtain Shannon's information, closing the gap in the range that the parameter can achieve. As Shannon's information is also nonnegative for discrete events, we get a complete nonnegative information for all possible values of our parameter.

4. Additivity

$$S_{\alpha} = \frac{1}{1-\alpha} \cdot \ln \left[\sum_{k=1}^{n} p_k^{\alpha} \right] \quad \textit{with} \quad \alpha \neq 1 \tag{3.12}$$

additionally has the following properties:

$S_\alpha(A)$ is symmetric, normalized, additive and nonnegative (decisive, expandable) [1].

The additivity of the information of two independent random variables can be obtained by computing the Rényi α-information of the joint distribution density of two independent random variables, which means that $f_{z,y} = f_x \cdot f_y$ the joint distribution in equal to the product of the marginal distribution densities. As the joint distribution density can be created from the product of the marginal distribution densities, we are able to transform the kernel of the integral and thus to separate the integral into to integrals, which means:

$$S_{\alpha Z} = S_{\alpha X} + S_{\alpha Y} \qquad \text{for independent random variables} \tag{3.13}$$

Additivity actually means that the entropy of the joint distribution density can be computed from the entropies of a marginal distribution and a conditional conditional, i.e. $f_{x,y} = f_{x|y} \cdot f_y$. If the random variables are independent, then this approach can be simplified and we again obtain the formulation we previously examined.

The entropy $S_\alpha(A)$ is strictly additive. Thus it is not easy to define a concept of a transformation in the form of differences of entropies.

3.3 Relation between Shannon's and Rényi's Information

$\ln x$ is a concave function and satisfies the Jensen inequality

$$\ln\left(\sum_{k=1}^{n} p_k \cdot x_k\right) \geq \sum_{k=1}^{n} p_k \cdot \ln x_k \tag{3.14}$$

Let $x_k = p_k^{\alpha-1}$ with $\alpha \neq 1$, which provides

$$\begin{aligned}\ln\left(\sum_{k=1}^{n} p_k \cdot p_k^{\alpha-1}\right) &\geq \sum_{k=1}^{n} p_k \cdot \ln p_k^{\alpha-1}\\ \ln\left(\sum_{k=1}^{n} p_k^{\alpha}\right) &\geq (\alpha-1) \cdot \sum_{k=1}^{n} p_k \cdot \ln p_k\\ \ln\left(\sum_{k=1}^{n} p_k^{\alpha}\right) &\geq -(1-\alpha) \cdot \sum_{k=1}^{n} p_k \cdot \ln p_k\end{aligned} \tag{3.15}$$

This presents the connection to Shannon's entropy, and we are able to proceed with further transformation with regard to α.

The relation of Rényi's entropy and Shannon's entropy depends on the parameter α. We may thus define an α- transformation

$$I_{T\alpha} = S(A) - S_\alpha(A) \tag{3.16}$$

and write:

$$\begin{aligned}I_{T\alpha} &= S(A) - S_\alpha(A) < 0 \quad for \quad 0 < \alpha < 1\\ I_{T\alpha} &= S(A) - S_\alpha(A) > 0 \quad for \quad \alpha > 1\end{aligned} \tag{3.17}$$

This α-transformation occurs if we measure the information of a certain event with Shannon's measure and with Rényi's measure of information and calculate the difference between the observe measures of information of the same event A.

Additionally, combining with previous properties we have, we get:

$$\begin{aligned} \ln n \geq S_\alpha(A) > S_1(A) \geq -\ln p_{max} \quad & for \;\; 0<\alpha<1 \\ \ln n \geq S_\alpha(A) = S_1(A) \geq -\ln p_{max} \quad & for \;\; \alpha = 1 \\ \ln n \geq S_1(A) > S_\alpha(A) \geq -\ln p_{max} \quad & for \;\; 1<\alpha \end{aligned} \tag{3.18}$$

This describes the connection between Rényi's entropy and Shannon's entropy for different values of the parameter α.

Chapter 4

Kullback's Measures of Information and Divergence

4.1 Kullback's Information

The information that Kullback himself preferred the name 'discrimination information' is generated from the difference of two Shannon information functions[34]. As a difference of these information functions, it is — contrary to Shannon's information — invariant with respect to a transformation of the coordinate system. For a continuous random variable, the result

$$I_{Kullback} = G_1 = \int_{\Xi} \ln\left(\frac{f(\xi)}{f_s(\xi)}\right) \cdot f(\xi) \cdot \mathrm{d}\xi \tag{4.1}$$

is Kullback's 'discrimination information'. Some easy rearrangements lead to:

$$G_1 = \int_{\Xi} \ln(f(\xi)) \cdot f(\xi) \cdot \mathrm{d}\xi - \int_{\Xi} \ln(f_s(\xi)) \cdot f(\xi) \cdot \mathrm{d}\xi \tag{4.2}$$

$$G_1 = -S(f(\xi)) + S\left(f_s(\xi), f(\xi)\right) \tag{4.3}$$

now we can immediately notice the relation to Shannon's information S. The mean of the information difference together with the application of Shannon's information led us to Kullback's information. This information arises when we replace a given probability distribution by another probability distribution.The derivation from the G-divergence demonstrates that the discrimination information describes the information difference between two distribution densities.

Furthermore Kullback's information allows us to describe the information of the random variable x contained in the random variable y

$$G_1 = \int_P \ln\left(\frac{f_{y,x}(\rho,\xi)}{f_x(\xi)\cdot f_y(\rho)}\right)\cdot f_{y,x}(\rho,\xi)\cdot \mathrm{d}\rho\cdot \mathrm{d}\xi \tag{4.4}$$

This information is also known as Kolmogorov's information (mutual information), and it is a special, symmetric case of Kullback's information.

4.2 Kullback-Leibler Divergence 1951

To measure the difference between two probability distributions over the same variable x, Kullback and Leibler(1951) were the first to introduce an information measure between two distribution density functions, the Kullback-Leibler divergence. The KL divergence, which is closely related to relative entropy, information divergence, and discrimination information(previous section), is a non-symmetric measure of the difference between two probability distributions p and q. Specifically, the Kullback-Leibler (KL) divergence of $q(x)$ from $p(x)$, denoted $D_{KL}(p,q)$, is a measure of the information lost when q is used to approximate p.

Let p and q are two probability distributions of a discrete random variable. That is, both p and q sum up to 1, and $p > 0$ and $q > 0$. $D_{KL}(p,q)$ is defined by

$$D_{KL}(P\|Q) = \sum_{i=1}^{n} p_i \cdot \ln\frac{p_i}{q_i} \qquad for\ all\ P,Q \in \Delta_n \tag{4.5}$$

Let's imagine an application: the KL divergence measures the expected number of extra bits required to code samples from p when using a code based on q, rather than using a code based on p. Typically p represents the "true" distribution of data, observations, or a precisely calculated theoretical distribution. The measure q typically represents a theory, model, description, or approximation of p.

The continuous version of the KL divergence is

$$D_{KL}(p(x)\|q(x)) = \int_{-\infty}^{\infty} p(x) \ln \frac{p(x)}{q(x)} \mathrm{d}x \tag{4.6}$$

Although the KL divergence measures the "distance" between two distributions, it is not a distance measure. This is because that the KL divergence is not a metric measure. It is not symmetric: the KL from $p(x)$ to $q(x)$ is generally not the same as the KL from $q(x)$ to $p(x)$. Furthermore, it need not satisfy triangular inequality. Nevertheless, $D_{KL}(P\|Q)$ is a non-negative measure. $D_{KL}(P\|Q) \geq 0$ and $D_{KL}(P\|Q) = 0$ if and only if $P = Q$.

4.3 Discrete and Continuous Measures of Entropy and Relative Entropies

Kullback's information measure is a relative information, which means that it describes the difference between two continuous distribution densities or between two probability distributions in discrete case. This relative definition of the information measure provides three advantages:

At first, the value of the calculated information is directly a measure, which describes the difference between amounts of information in the two random variables. We there fore do not require any additional generation of a reference value in the form of a second information. On the other hand it is a disadvantage of this information that we cannot directly determining which random variable contains more information. This is caused by the nonnegative of this measure.

The great advantage of this measure is that it does not change with a change of the coordinate system. Thus we can use this measure to determine the information differences before and after some arbitrary transformations. Furthermore, this measure allows us to compare discrete and continuous entropies. We already know from Shannonn's information that the transformation from discrete to continuous systems leads to an additional term in the information. However, differences between two Rényi entropies can be compared, because this additional term vanishes as soon as we build difference.

Kullback's information offers the advantages that we directly obtain a difference information (absolute information measures do not directly provide any meaning because we have neither a reference scale nor a unit element), and that we may compare discrete and continuous entropies. Additionally, this information measure is invariant regarding transformations. This advantage is a general property of the relative information measures, calculating a comparison between two distribution densities or between two probability distributions.

In the book [30], J.N. Kapur argues that

> We are more fortunate for the measure of directed divergence, since here the limit of measures for discrete probability distributions give measures for continuous variate distribution. And it is therefore desirable to consider the concept of directed divergence as more basic than that of entropy and to express the concept of entropy in terms of the concept of directed divergence.

So the Kullback-Leibler Divergence turns out to be a critical part of our research. Especially at the beginning of our research, we need to build our intuitive models from basic measures of information. Detail discussions and conclusions can be found in Chapter 6.

Chapter 5

Summary of Information, Entropy, and Divergence Measures

5.1 Measures of Information

We have now presented several information measures and describe their properties more from the descriptive point of view. These information functions are the most familiar ones, which are commonly used in applications. It is impossible to enumerate all information functions, because every stated problem leads to the generation of another information function. The information functions introduced demonstrate already the most important properties and possible variances that are typical for such information functions.

Measures of information and measures of entropy are closely related to each other, which caused our excursion into thermodynamics. There the concept of entropy arose in the description of thermal energy and it is nowadays used in the statistical description of physical state. The entropy, as it is described in physics, describes the probability of the occupation of certain states. It is a measure of uncertainty, which is exactly the concept we used in the documentation Shannon's information and our current research. Information is a concept referring to the future and describing the possibility of a gain of information by an observation. So we can also denote the entropy functions as information functions, because the only difference is the reference we used in the interpretation.

The examination of information cannot be restricted to one optimal information function initially, because we cannot always foresee which information is the best for the calculations. Cambel [11] states this fact this way:

> There is no one best definition. Use the one best suited for your purpose. Also, please do not overlook the possibility that describing certain systems may require evaluating several of its entropy functions.

It is pretty clear in the examination of maximum entropy principle where we need to generate entropy functions for arbitrary arrangements.

By considering several different descriptions of information, however, we did not get a detailed definition for the concept of information. On the basis of certain demands that we required the information to meet, we found some measures that behave according out intuition.

Here is a table present the most familiar information functions and show their connections. Some of information measures are mentioned in previous chapters, others will be briefly introduced in following sections of this chapter.

		POSTULATES		VARIED POSTULATES
	$\swarrow$	$\downarrow$		$\downarrow$
S_α Rényi's α-information	$\rightarrow$ ΔS_α	Rényi's G_α-information		Daroczy's entropy
$\downarrow$ $\alpha \rightarrow 1$		$\downarrow$ $\alpha \rightarrow 1$	$\searrow$ $f_{yx}/$ $f_x f_y$	Generalized mutual information
Shannon's S-information	$\rightarrow$ ΔS	Kullback's G_1-information	$\rightarrow$ ΔG_1	Kullback's D-divergence
		$\downarrow$ Series expansion Differential quotient	$\searrow$ $f_{yx/}$ $f_x f_y$	
		Fisher's information matrix		Kolmogorov's Transformation

Table 5.1: Connections between the information measures

In the following sections. we present only an enumeration of generalized measures, without any detailed examination of the presented measures, because this would beyond the cope of this thesis.

5.2 Entropy Measures

The most important properties of such entropy measures, which are met by Shannon's information, are:

1. *Additivity*

$$H(P \cdot Q) = H(P) + H(Q) \tag{5.1}$$

   For $P = \{p_1, p_2, ..., p_n\} \in \Delta_n^0$ and $Q = \{q_1, q_2, ..., q_m\} \in \Delta_m^0$ and for the product space $P \cdot Q = \{p_1q_1, p_1q_2, ..., p_1q_m, p_2q_1, p_2q_2, ..., p_nq_m\} \in \Delta_{nm^0}$ where $P = \{p_1, p_2, ..., p_n\} \in \Delta_n^0 =$ set of all probabilities of n events, where certain events may occur with a probability $p_i = 0$.

2. *Recursivity*

$$H(p_1, p_2, ..., p_n) = H(p_1 + p_2, p_3, ..., p_n) = (p_1 + p_2) \cdot H\left(\frac{p_1}{p_1 + p_2}, \frac{p_2}{p_1 + p_2}\right) \tag{5.2}$$

   $p_1 + p_2 > 0\ for\ all\ P = \{p_1, p_2, ..., p_n\} \in \Delta_n^0$

3. *Summation property*

$$H(P) = \sum_{i=1}^{n} f(p_i) \qquad P = p_1, p_2, ..., p_n \in \Delta_n^0 \tag{5.3}$$

   for Shannon's measure we get $f(p) = p \ln p$.

For all $P = \{p_1, p_2, ..., p_n\} \in \Delta_n^0 =$ set of all probabilities of n events, where certain events may occur with probability $p_i = 0$, there exist several entropy(information) measures:

Measure $\mathcal{I} - 1$ Shannon(1948)

$$\mathcal{I}_1 = -\sum_{i=1}^{n} p_i \cdot \ln p_i \tag{5.4}$$

Measure $\mathcal{I} - 2$ Rényi(1961)

$$\mathcal{I}_2 = \frac{1}{1 - \alpha} \cdot \ln\left(\sum_{i=1}^{n} p_i^{\alpha}\right) \quad \alpha \neq 1,\ \alpha > 0 \tag{5.5}$$

Measure $\mathcal{I}-3$ Acél and Daróczy (1963)

$$\mathcal{I}_3 = -\frac{\sum_{i=1}^{n} p_i^{\alpha} \cdot \ln p_i}{\sum_{i=1}^{n} p_i^{\alpha}} \qquad \alpha > 0 \tag{5.6}$$

Measure $\mathcal{I}-4$

$$\mathcal{I}_4 = \frac{1}{s-\alpha} \cdot \ln \left(\frac{\sum_{i=1}^{n} p_i^{\alpha}}{\sum_{i=1}^{n} p_i^{s}} \right) \qquad \alpha \neq s,\ \alpha \neq 1,\ \alpha > 0 \quad (\textit{see } \mathcal{I}_2) \tag{5.7}$$

Measure $\mathcal{I}-5$

$$\mathcal{I}_5 = \frac{1}{s} \cdot \arctan \left(\frac{\sum_{i=1}^{n} p_i^{\alpha} \cdot \sin(s \cdot \ln p_i)}{\sum_{i=1}^{n} p_i^{\alpha} \cdot \cos(s \cdot \ln p_i)} \right) \qquad s \neq 0,\ \alpha > 0 \tag{5.8}$$

Measure $\mathcal{I}-6$ Varma (1966)

$$\mathcal{I}_6 = \frac{1}{m-\alpha} \cdot \ln \left(\sum_{i=1}^{n} p_i^{\alpha-m+1} \right) \qquad m-1 < \alpha < m, m \geq 1 \tag{5.9}$$

Measure $\mathcal{I}-7$

$$\mathcal{I}_7 = \frac{1}{m \cdot (m-\alpha)} \cdot \ln \left(\sum_{i=1}^{n} p_i^{\frac{\alpha}{m}} \right) \qquad m-1 < \alpha < m, m \geq 1 \tag{5.10}$$

Measure $\mathcal{I}-8$ Kapur (1967)

$$\mathcal{I}_8 = \frac{1}{1-t} \cdot \ln \left(\frac{\sum_{i=1}^{n} p_i^{t+s+1}}{\sum_{i=1}^{n} p_i^{s}} \right) \qquad t \neq 1,\ t > 0, s \geq 1 \quad (\textit{see } \mathcal{I}_6) \tag{5.11}$$

Measure $\mathcal{I}-9$ Havrda and Charvát (1967)

$$\mathcal{I}_9 = \frac{1}{2^{1-s}-1} \cdot \left(\sum_{i=1}^{n} p_i^{s} - 1 \right) \qquad s \neq 1, s > 0 \tag{5.12}$$

Measure $\mathcal{I}-10$ Belis and Guiasu (1968)

$$\mathcal{I}_{10}=-\frac{\sum_{i=1}^{n}p_i\cdot w_i\cdot\ln p_i}{\sum_{i=1}^{n}p_i\cdot w_i}\quad w_i>0,\ i=1,2,3...,n\quad (\textit{see } \mathcal{I}_3)\tag{5.13}$$

Measure $\mathcal{I}-11$ Rathie (1970)

$$\mathcal{I}_{11}=\frac{1}{1-\alpha}\cdot\ln\left(\frac{\sum_{i=1}^{n}p_i^{\alpha+s_i+1}}{\sum_{i=1}^{n}p_i^{s_i}}\right)\quad \alpha\neq 1,\alpha>0,\quad s_i\geq 1,\ i=1,2,3,...n\ (\textit{see } \mathcal{I}_8)\tag{5.14}$$

Measure $\mathcal{I}-12$ Arimoto (1971)

$$\mathcal{I}_{12}=\frac{1}{2^{t-1}-1}\cdot\left(\left(\sum_{i=1}^{n}p_i^{\frac{1}{t}}\right)^{t}-1\right)\quad t\neq 1,\ t>0\tag{5.15}$$

Measure $\mathcal{I}-13$ Sharma and Mittal (1975)

$$\mathcal{I}_{13}=\frac{1}{2^{1-s}-1}\cdot\left[\exp_2\left((s-1)\cdot\sum_{i=1}^{n}p_i\cdot\ln p_i\right)-1\right]\quad s\neq 1,s>0\tag{5.16}$$

Measure $\mathcal{I}-14$

$$\mathcal{I}_{14}=\frac{1}{2^{1-s}-1}\cdot\left[\left(\sum_{i=1}^{n}p_i^{\alpha}\right)^{\frac{s-1}{\alpha-1}}-1\right]\quad s\neq 1,s>0\tag{5.17}$$

Measure $\mathcal{I}-15$ Taneja (1975)

$$\mathcal{I}_{15}=-2^{\alpha-1}\cdot\sum_{i=1}^{n}p_i^{\alpha}\cdot\ln p_i\quad \alpha>0\tag{5.18}$$

Measure $\mathcal{I}-16$

$$\mathcal{I}_{16}=\frac{1}{2^{1-\alpha}-2^{1-s}}\cdot\sum_{i=1}^{n}(p_i^{\alpha}-p_i^{s})\quad \alpha\neq s,\alpha>0,s>0\tag{5.19}$$

Measure $\mathcal{I}-17$

$$\mathcal{I}_{17} = -\frac{2^{\alpha-1}}{\sin(s)} \cdot \sum_{i=1}^{n} p_i^{\alpha} \cdot \sin(s \cdot \ln p_i) \quad s \neq k\pi, k = 0,1,2,...\alpha > 0 \tag{5.20}$$

Measure $\mathcal{I}-18$ Picard (1979)

$$\mathcal{I}_{18} = -\frac{\sum_{i=1}^{n} v_i \cdot \ln p_i}{\sum_{i=1}^{n} v_i} \qquad (see\ \mathcal{I}_3) \tag{5.21}$$

Measure $\mathcal{I}-19$

$$\mathcal{I}_{19} = \frac{1}{1-\alpha} \cdot \ln\left(\frac{\sum_{i=1}^{n} p_i^{\alpha-1} \cdot v_i}{\sum_{i=1}^{n} v_i}\right) \qquad \alpha \neq 1, \alpha > 0 (see\ \mathcal{I}_3) \tag{5.22}$$

Measure $\mathcal{I}-20$

$$\mathcal{I}_{20} = \frac{1}{2^{1-s}-1} \cdot \left[\frac{\exp_2\left((s-1) \cdot \sum_{i=1}^{n} v_i \cdot \ln p_i\right)}{\sum_{i=1}^{n} v_i} - 1\right] \qquad s \neq 1, s > 0 (see\ \mathcal{I}_1 3) \tag{5.23}$$

Measure $\mathcal{I}-21$

$$\mathcal{I}_{21} = \frac{1}{2^{1-s}-1} \cdot \left[\left(\frac{\sum_{i=1}^{n} v_i \cdot p_i^{\alpha-1}}{\sum_{i=1}^{n} v_i}\right)^{\frac{s-1}{\alpha-1}} - 1\right] \quad s \neq 1, s > 0, \alpha > 0\ (see\ \mathcal{I}_1 4) \tag{5.24}$$

These functions exchanging the expectation over p_i by the expectation over v_i for all entropies. The weights are $v_i > 0\ for\ i = 1,2,...,n$ and $P = \{p_1, p_2, ..., p_n\} \in \Delta_n$, i.e. $p_i = 0$ is not allowed.

Table 5.2: Entropy Measures

Taneja uses this entropies to generate a *unifies*(α, s)-*entropy* with

$$E_\alpha^s(P) = \begin{cases} H_\alpha^s(P) & for\ \alpha \neq 1, s \neq 1, \alpha > 0 \\ H_1^s(P) & for\ \alpha = 1, s \neq 1 \\ H_\alpha^1(P) & for\ \alpha \neq 1, s = 1, \alpha > 0 \\ H(P) & for\ \alpha = 1, s = 1 \end{cases}$$

and this most general form can be examined in detail.

During this examination, Taneja defines certain properties which he uses for the description of analytics and algebraic properties of the unified (α, s)-entropy. The complete explanation of this unified entropy can be found in [48].

5.3 Information, Entropy and Probability

5.3.1 How Can We Describe Information?

Of course, like we mentioned before, we have not been able to answer these questions completely. Certain fundamental questions remain, which we may never be able to answer in a satisfying way. The nature of information can only be partially described by the objective descriptions that we applied, while other aspects of information are less amenable to scientific descriptions. The fundamental idea is very old and certainly ‘rethought’ in many different ways, but no closed definition has been found up to the present. Nevertheless, it seems that information and its basic meaning is not only based on physically measurable quantities. Information is able to appear in different quantities and is able to change carrier easily, which demonstrates a certain independence of the measurable quantities, which we are used to.

However, we are able to describe and apply a certain part of the information, the syntactic information, with the information functions described. The definition of the whole concept of information may be impossible for us, so that we have to rely on our familiar methods. Though most popular applicable concept of information is based on a limited definition, the resulting applications are very extensive and the information age, which is now propagated, promise further applications of both information processing and transmission. Like we mentioned in the Chapter

1 and always emphasize in our research that the "information transmission" are finally not responsible for the meaning or the sense of the transmitted messages. This responsibility is left to us. So it is up to us to extract the reasonable information out of the growing flood of information with modern techniques such as Big Data, Machine Learning, etc..

The concept of entropy as well as the concept of information are closely related to the probability of the events. The information functions are the expectation values of distribution densities, which describes the probabilities of the possibilities of the possible events. These probabilities, which we assign to the events, have to be defined in someway before we use them. One possibility of a definition is the use of the relative frequency

$$h(A) = \frac{n(A)}{A} \tag{5.25}$$

describes the number of the elements A occurring in n samples. The probability can be understood as the limit of this relative frequency for large n (numbers of samples) or as the expectation of the relative frequency. Both intuitive definitions of the probability have certain disadvantages, prohibiting a direct mathematical application of these heuristic denotations. The determination of the probability as the expectation of the relative frequencies already requires the definition of the probability in the computation of the expectation value $P(A) = E\{h(A)\}$ and it thus leads to a circular argument, which is not allowed in a correct definition. The second approach computes the limit

$$P(A) = \lim_{n\to\infty} h(A) = \lim_{n\to\infty} \frac{n(A)}{n} \tag{5.26}$$

These limit cannot be computed with the common analytic methods and the observation of an infinite number of samples is not realistic. The more samples we observe, the more the relative frequency should approach the theoretical probability.

5.3.2 Characterizations toward the Quantitative Aspects in Our Research

Now let's recall those quantitative aspects of the main framework of our research introduced in Section 1.1 that we suggested: *Question difficulty, Answer Depth and Knowledge Structure.*

Theorem 5.3.1. *Let the functional $G(\Omega, Q, P)$ where $Q = \{C_1, ..., C_r\}$ satisfy Postulates 1* *through 6 (see [39]). Then it has the form*

$$G(\Omega, Q, P) = \frac{\displaystyle\sum_{j=1}^{r} u(C_j)P(C_j)\log\frac{1}{P(C_j)}}{\displaystyle\sum_{j=1}^{r} P(C_j)}$$

where $u(C_j) = \frac{\int_{C_j} u(\omega)dP(\omega)}{P(C_j)}$ *and* $u : \Omega \to \mathbb{R}$ *is am integrable nonnegative function on the parameter space* Ω.

Theorem 5.3.2. *The answer depth functional* $Y(\Omega, Q, P, V(Q))$ *has the form*

$$Y(\Omega, Q, P, V(Q)) = \sum_{k=1}^{m} Pr(V(Q) = s_k)\frac{\displaystyle\sum_{j=1}^{r} u(C_j)P^k(C_j)\log\frac{P^k(C_j)}{P(C_j)}}{\displaystyle\sum_{j=1}^{r} P^k(C_j)},$$

where $P^k \equiv P^{V(Q)=s_k}$ *is the measure on* Ω *conditioned on reception of* $V(Q) = s_k$ *and* $u(C_j) = \frac{1}{P(C_j)}\int_{C_j} u(\omega)dP(\omega)$ *and the function* $u : \Omega \to \mathbb{R}$ *is the same function that is used in the question difficulty functional* $G(\Omega, Q, P)$.

Things about Knowledge Structure is more complicated and from previous sections about knowledge and probability, we know that describing one's Knowledge by Probabilities could be a proper direction. Currently we are working on the measure of Knowledge Structure so deeply analyzing of knowledge are omitted here.

From previous introduction, we know that the measure of information and the measure of entropy are closed to each other. And according to the definition of information in our research, it is very suitable to measure these quantitative aspects of information by the measures based on Shannon's Information and Entropy.

From Chapter 2-5, we know that information measures based on Shannon's Entropy may or may not have these characterizations like **Symmetry, Normality, Expansibility, Decisivity, Branching, Additivity, Recursivity, Summation**. Furthermore, the most important and most widely used is additivity, resursivity, and summation. Refer to Table 5.2 that lists several entropy/information measures, some special cases needs to be focused so that we can distinguish the best measures of information for our research.

Rényi's entropy H_α is an extension with a generalized mean

$$H_\alpha(P) = \frac{1}{1-\alpha} \cdot \ln\left(\sum_{i=1}^{n} p_i^\alpha\right) \tag{5.27}$$

It is an entropy of the order α. It meets the additivity, but it cannot be calculated with a recursive formulation and even the summation property does not hold. Havrda and Charvát introduced the entropy H^s of grade s

$$H^s(P) = \frac{1}{2^{1-s}-1} \cdot \left(\sum_{i=1}^{n} p_i^s - 1\right) \tag{5.28}$$

which does not meet the additivity, but is recursive in the grade s

$$H^s(p_1, p_2, ..., p_n) = H^s(p_1 + p_2, p_3, ..., p_n) = (p_1 + p_2)^s \cdot H^s\left(\frac{p_1}{p_1 + p_2}, \frac{p_2}{p_1 + p_2}\right) \tag{5.29}$$

$p_1 + p_2 > 0\ for\ all\ P = \{p_1, p_2, ..., p_n\} \in \Delta_n^0$ and even the sum representation $H^s(P) = \sum_{i=1}^{n} f(p_i)$ is possible. Moreover, there is third index, indicating an entropy ${}_tH$ of the kind t, and it has been presented by Arimoto with the entropy

$${}_tH(p) = \frac{1}{2^{t-1}-1} \cdot \left(\left(\sum_{i=1}^{n} p_i^{\frac{1}{t}}\right)^t - 1\right) \tag{5.30}$$

If we replace $i/t\ by\ \alpha$, we may easily derive Rényi's entropy.

Sharma and Mittai presented an entropy H_α^s of order α and grade s

$$H_\alpha^s(P) = \frac{1}{2^{1-s}-1} \cdot \left[\left(\sum_{(} i = 1)^n p_i^\alpha\right)^{\frac{s-1}{\alpha-1}} - 1\right] \tag{5.31}$$

It is neither additive nor recursive, and they do not even meet the sum-representation. They are generalizations of entropies presented so far and can be reduced to the special forms.

Information measures like all these measures mentioned above may not be suitable for our research, according to the requirement of the definition functions in Theorem 6.1.3 and 6.1.2. Not only because that these measures doesn't meet certain characterizations like additivity or recursivity, since J. N. Kapur argues in the book [30]

$$RECURSIVITY \equiv BRANCHING \equiv\ ADDITIVITY \equiv\ DECOMPOSABILITY$$

- Recursivity implies that entropy measure for n outcomes can be found when entropy measures for m outcomes are known where $m = 1, 2, ..., n - 1$.
- Branching implies that entropy for $p_1, p_2, ..., p_n$ can be expressed in terms of entropy for $p_1 + p_2, p_3, ..., p_n$ and entropy of p_1, p_2.
- Additivity means that entropy of a joint distribution can be expressed as the sum of two entropies, one of which is the entropy of one marginal distribution and the other is the conditional distribution in the second distribution.
- Decomposability emphasizes that the total entropy can be decomposed into entropy within classes and entropy between entropies.

but also that certain measures of information are established on their own applications with unique description of information. Some of them are not suitable for our research.

So in conclusion, based on our research and all the relative introduction in this thesis, we have conclude that the **Kullback-Leibler Divergence** (also known as Kullback's information, relative information etc) is a proper measure of information for our research based on our definition of information here (see chapter 6).

Kullback-Leible Divergence do not have the characterization of Symmetric, but it has the property of additivity and recurisivity. Formulations like

$$D_{KL}(P\|Q) = \sum_{i=1}^{n} p_i \cdot \ln \frac{p_i}{q_i} \qquad for\ all\ P, Q \in \Delta_n \tag{5.32}$$

with its definition of description of difference between two informations quite match not only the quantitative aspects we suggested but also meet the whole definition of information we derived for the general purpose of our research.

5.4 Relative Information and Divergence

5.4.1 On the Information Gain

Now after we define the knowledge of each characters, the next question would be how to describe the knowledge(information) change process. Explicitly, how to the describe and measure the information between the agent and source. And most importantly, how to measure the information the the agent gain.

The basic consideration is connected with the two terms 'gain of information' and 'gain of uncertainty', which differ only by a negative sign. In principle, we have two possible ways to define the Information Acquisition. The first way is to create mean of all single gains of information. The second way is to create the mean of all single uncertainty differences and multiply the results by -1. In the case of Shannon's information, both approach lead to the same result. First, we generate the mean value of a function of the changed probability $\frac{q_k}{p_k} = \frac{P(A_k|B_j)}{P(A_k)}$ and called the result a gain of information. In the second formulation we calculate the mean value of the same function, but use the reciprocal argument $\frac{q_k}{p_k} = \frac{P(A_k|B_j)}{P(A_k)}$. This is equivalent to a calculation of the mean of uncertainty and we obtain the gain of information simply by multiplying the calculated mean value by the factor -1. The results is same in both approaches, when we use Shannon's information.

This identity of both approaches does not hold if we apply Rényi's information measure of the order α. There we obtain the information measurement by the equation

$$\mathcal{I}_2 = I_{\text{Rényi}} = \frac{1}{1-\alpha} \cdot \ln\left(\sum_{k=1}^{n} p_k \cdot p_k^{\alpha-1}\right) \quad \alpha \neq 1,\ \alpha > 0 \tag{5.33}$$

Here we compute the expectation value of $p_k^{\alpha-1}$. Replacing the single probabilities by the quotient $\frac{q_k}{p_k} = \frac{P(A_k|B_j)}{P(A_k)}$, leads us to the gain of information in the formulation of Rényi's information.

$$I_\alpha(Q\|P) = \frac{1}{\alpha-1} \cdot \ln\left[\sum_{k=1}^{n} \frac{q_k^\alpha}{p_k^{\alpha-1}}\right] \tag{5.34}$$

as Rényi's gain of information for replacing the given discrete probability distribution $P =$

$(p_1, p_2, ..., p_n)$ by the probability Q. The second possibility is to replace the quotient $\dfrac{q_k}{p_k} = \dfrac{P(A_k|B_j)}{P(A_k)}$ by its reciprocal quotient $\dfrac{q_k}{p_k} = \dfrac{P(A_k|B_j)}{P(A_k)}$, which leads to a measure of uncertainty U_α. This provides the equation

$$U_\alpha(Q\|P) = \frac{1}{1-\alpha} \cdot \ln\left[\sum_{k=1}^{n} \frac{q_k^{2-\alpha}}{p_k^{1-\alpha}}\right] \tag{5.35}$$

The information difference always vanishes for $\alpha = 2$ and therefor this formulation is not suitable for denoting an information difference. However, the other formulation of the information difference merely consists of the special case $\alpha = 1$.

If we apply Jensen's inequality to the gains of information$I_\alpha(Q\|P)$, $I_1(Q\|P)$ with the concave logarithm function and another function x^α, which is concave in the interval $0 < \alpha < 1$,but convex for $\alpha > 1$, we find

$$I_\alpha(Q\|P) \geq 0$$

and

$$I_1(Q\|P) \geq 0$$

(please see Kullback's information in Chapter 4). The gain of information is thus always positive and it only equals to zero when the two probabilities p and q are equal.

Rényi's gain of information, however, is not a symmetric function of the two arguments p_k and q_k, because the resulting gains of information are in fact nonnegative, but in general they do not provide the same gains of information. To obtain a symmetric gain of information, independent of the order of replacement, it is possible to define a J-divergence. Considering that in our research, it is more about to analyze the information acquisition from an agent for knowledge change, symmetric property is not that important in our case, I think. So detail about J-divergence is omitted here. More useful information can be found [48].

For the special case that $Q = \{q_1, q_2, ..., q_n\} = \{1/n, 1/n, ..., 1/n\}$ is a uniform probability distribution and that we replace it by another probability distribution P

$$I_\alpha(Q\|P) = \frac{1}{\alpha-1} \cdot \ln\left[\sum_{k=1}^{n} \frac{q_k^{\alpha}}{p_k^{\alpha-1}}\right] = \frac{1}{\alpha-1} \cdot \ln\left[\sum_{k=1}^{n} n^{\alpha-1} \cdot p_k^{\alpha}\right]$$

$$I_\alpha(Q\|P) = \ln n - \frac{1}{\alpha - 1} \cdot \ln \left[\sum_{k=1}^{n} p_k^\alpha \right]$$

$$I_\alpha(P\|Q) = I_{Hartley}(Q) - I_\alpha(P)$$

$$I_\alpha(P\|Q) = I_{Shannon}(Q) - I_\alpha(P) = I_1(Q) - I_\alpha(P)$$

There the information difference is equal to the difference of two information values. The uniform distribution density maximizes the information function and thus the gain of information is always positive. Thus the entropy described by the function decrease when we replace the uniform distribution density by any other distribution density.

This approach allows us to define a gain of information on the basis of completely missing a priori information, because the distribution density describing the complete lack of any previous information is the uniform distribution density. This is connection to Laplace's principle of insufficient reasoning, expressing the fact that we must not prefer any event, if we do not have priori information. We thus must have to assign equal probabilities to all events. It is quite important for us to use in our research that it provides an limit of description of agents' knowledge in a special case. If an agent know nothing about this objective, we should assign a uniform distribution to it and then calculate the information change after the agent receives the answer from the source.

5.4.2 The Key: Kullback-Leibler Divergence

In Chapter 4, we discuss that Kullback and Leibler introduced an information measure between two distribution density functions, the Kullbacl-Leibler divergence, discrimination function, relative information, directed divergence or Kullback's information.

$$D_{KL}(P\|Q) = \sum_{i=1}^{n} p_i \cdot \ln \frac{p_i}{q_i} \qquad for\ all\ P, Q \in \Delta_n \tag{5.36}$$

It is also shown that this divergence can be extended by applying a generalized mean, which has been realized by Rényi(1961)

$$D_\alpha^1(P\|U) = \frac{1}{(1-\alpha)} \cdot \ln \left(\sum_{i=1}^{n} p_i^\alpha \cdot u_i^{1-\alpha} \right) \quad \alpha \neq 1, \alpha > 0\ for\ all\ P, U \in \Delta_n \tag{5.37}$$

This divergence can be extended by applying a generalized mean, which has been realized by Rényi(1961), who

$$D_\alpha^1(P\|U) = \frac{1}{(1-\alpha)} \cdot \ln\left(\sum_{i=1}^{n} p_i^\alpha \cdot u_i^{1-\alpha}\right) \qquad \alpha \neq 1, \alpha > 0 \; for \; all \; P, U \in \Delta_n \tag{5.38}$$

Another generation of the divergence is

$$D_s^s(P\|U) = \frac{1}{1-2^{1-s}} \cdot \left(\sum_{i=1}^{n} p_i^s \cdot u_i^{1-s} - 1\right) \qquad s \neq 1, s > 0, \; for \; all \; P, U \; \in \Delta_n \tag{5.39}$$

Both generalization can be reduced to the Kullback-Leibler distance by calculation of the limit for

$$\lim_{\alpha \to 1} D_\alpha^1(P\|U) = \lim_{s \to 1} D_s^s(D\|U) = D(P\|U) \tag{5.40}$$

Like previous section about entropy measure, all these generalized divergence measures can then be summarized by a unified directed divergence, defined by

$$D_\alpha^s(P\|U) = \begin{cases} D_\alpha^s(P\|U) & for \; \alpha \neq 1, s \neq 1, \alpha > 0 \\ D_1^s(P\|U) & for \; \alpha = 1, s \neq 1 \\ D_\alpha^1(P\|U) & for \; \alpha \neq 1, s = 1, \alpha > 0 \\ D(P\|U) & for \; \alpha = 1, s = 1 \end{cases}$$

for all $P, U \in \Delta_n$.

To circumvent difficulties, which may occur, when the probabilities are equal to zero, the probability space is assumed to be given by Δ_n, so that zero probabilities do not need further attention in the definition.

For all $P, U \in \Delta_n$ the unified (α, s)-directed divergence $D_\alpha^s(P\|U)$ has the following properties:

(i) $D_\alpha^s(P\|U) \geq 0$ for all $\alpha > 0$ and all s.

(ii) $D_\alpha^s(P\|U)$ is a convex function of the pair $(P, U) \in \Delta_n \times \Delta_n$ for all $s \geq \alpha > 0$

(iii) $D_\alpha^s(P\|U)$ is an increasing functions of α (for constant s)

(iv)

$$D_\alpha^s\left(\sum_{i=1}^{\sigma} p_i, 1-\sum_{i=1}^{\sigma} p_i \Big\| \sum_{i=1}^{\sigma} u_i, 1-\sum_{i=1}^{\sigma} u_i\right) \leq D_\alpha^s(P\|U)$$

$$\leq D_\alpha^s\left(p_1, p_2, ..., p_\sigma, \frac{1-\sum_{i=1}^{\sigma} p_i}{n-1}, ..., \frac{1-\sum_{i=1}^{\sigma} p_i}{n-1} \Bigg\| u_1, u_2, ..., u_\sigma, \frac{1-\sum_{i=1}^{\sigma} u_i}{n-1}, ..., \frac{1-\sum_{i=1}^{\sigma} u_i}{n-1}\right)$$

$$for \ 1 \leq \sigma < n \tag{5.41}$$

(v) Let

$$P(c) = \left(\sum_{i=1}^{n} p_i \cdot c_{i1}, \sum_{i=1}^{n} p_i \cdot c_{i2}, ..., \sum_{i=1}^{n} p_i \cdot c_{in}\right) \in \Delta_n$$

$$U(c) = \left(\sum_{i=1}^{n} u_i \cdot c_{i1}, \sum_{i=1}^{n} u_i \cdot c_{i2}, ..., \sum_{i=1}^{n} u_i \cdot c_{in}\right) \in \Delta_n$$

with

$$\sum_{i=1}^{n} c_{ik} = \sum_{k=1}^{n} c_{ik} = 1 \qquad c_{ik} \geq 0, \ for \ i, k = 1, 2, ..., n$$

Then we get

$$D_\alpha^s(P(c)\|U(c)) \leq D_\alpha^s(P\|U) \tag{5.42}$$

(vi)

$$D_\alpha^s(P\|U) = \begin{cases} \leq D_\alpha^1(P\|U) & for \ \ s < 1 \\ \geq D_\alpha^1(P\|U) & for \ \ s > 1 \end{cases}$$

$$D_1^s(P\|U) = \begin{cases} \leq D(P\|U) & for \ \ s < 1 \\ \geq D(P\|U) & for \ \ s > 1 \end{cases}$$

$$D_\alpha^s(P\|U) = \begin{cases} \leq D_1^s(P\|U) & for \ \ 0 < \alpha < 1 \\ \geq D_1^s(P\|U) & for \ \ \alpha > 1 \end{cases}$$

$$D_\alpha^1(P\|U) = \begin{cases} \leq D(P\|U) & for \ \ 0 < \alpha < 1 \\ \geq D(P\|U) & for \ \ \alpha > 1 \end{cases}$$

There are also other kinds of divergence in the filed of Information measures, like Jensen's measures of divergence difference and J- divergence of Kullback and Leible. However, they are out of the scope of our research. For detailed information, please refer to [48].

It looks like this **Kullback-Leibler Divergence** can be a good measure for us to use in our research to measure the knowledge change of the agent after the agent receives the answer of the question he asks. It should be like this:

$$D(P_A \| P) = I(Q_{uestion}, A_{nswer}) \tag{5.43}$$

Detailed Information about why Kullback-Leibler Divergence is the *only* measure that satisfies our requirements can be shown in next chapter.

Chapter 6

Information Geometry and Information Acquisition Optimization

The field of Information Geometry explores geometric properties of manifolds of probability distributions and related manifolds. This approach proved to be rather fruitful to the fields of computational vision, optimization, signal processing and neural networks. A key concept of information geometry is a notion of a divergence function which can be thought of as generalized distance (not necessarily symmetric) between two point on a manifold [2, 3, 4]. In particular, two classes of divergence functions – Bregman divergences and invariant divergences play a particular role and help shed light on our problem – that of the proper quantitative measure of the degree of change from the original belief to the updated one.

6.1 Information Geometry of Divergence Functions

Given two points P and Q in space S, we may define a divergence $D[P : Q]$ which measures their discrepancy. The standard distance is indeed such a measure. However, there are many other measures frequently used in many areas of applications. In particular, for two probability distributions $p(x)$ and $q(x)$, one can define various measures $D[p(x) : q(x)]$ such as the Kullback-Leibler divergence and the Hellinger distance. A divergence is not necessarily symmetric, that is, the relation $D[P : Q] = D[Q : P]$ do not generally hold nor does it satisfy the triangular inequality. It usually has the dimension of squared distance, and a Pythagorean-like relation holds in some cases.

There are two typical classes of divergences: one is the class of **Bregman divergences** [10], introduced trough a convex function. The other is the class of **invariant dvergences**, called $f-$ divergence, where f is a convex function.

Bregman divergences are derived from convex functions. The Bregman divergence induces a dual structure through the Legendré transformation. It gives a geometrical structure consisting of a Riemannian metric and dually flat affine connections, called the dually flat Riemannian structure. A dually flat Riemannian manifold is a generalization of the Euclidean space, in which the generalized Pythagorean theorem and projection theorem hold. These two theorems provide powerful tools for solving problems in optimization, statistical inference and signal processing. We show that the Bregman type divergence is automatically induced from the dual flatness of a Riemannian manifold.

Then we study the class of invariant divergences . The invariance requirement comes from information monotonicity, which states that a divergence measure does not increase by coarse graining of information. This leads to the class of f-divergences. The α-divergences are typical examples belonging to this class, which also includes the Kullback-Leibler divergence as a special case. This class of divergences induces an invariant Riemannian metric given by the Fisher information matrix and a pair of invariant dual affine connections, the $\pm\alpha-$connections, which are not necessarily flat.

6.1.1 Bregman Divergence

Let $k_(z)$ be a strictly convex differentiable function defined in a space S with a local coordinate system z. Then, for two points z and y in S, we can define the following function

$$D[z:y] = k(z) - k(y) - \text{Grad } k(y) \cdot (z-y). \tag{6.1}$$

where, Grad k is the gradient vector

$$\text{Grad } k(z) = (\partial k(z)/\partial_{z_i}). \tag{6.2}$$

and the operator '·' denotes the inner product

$$\mathrm{Grad}k(y) \cdot (z-y) = \sum_i \frac{\partial k}{\partial_{y_i}}(z_i - y_i). \tag{6.3}$$

The function $D[z:y]$ satisfies the following condition for divergences:

1) $D[z:y] \geq 0$,

2) $D[z:y] = 0$ when and only when $= y$,

3) for small dz, Taylor expansion

$$D[z+dz:z] \approx \frac{1}{2}\sum g_{ij} dz_i dz_j \tag{6.4}$$

gives a positive-definite quadratic form.

We call $D[z:y]$ the Bregman divergence between two points z and y. In general, the divergence is not symmetric with respect to z and y so that

$$D[y:z] \neq D[z:y]. \tag{6.5}$$

There are some important theorems about the Bregman divergence and Riemannian metric. Some details and proofs given by S. Amari are omitted here and can be found in [2, 3, 4].

Theorem 6.1.1. *The Riemannian metrics g_{ij} and g^*_{ij} in their matrix form are mutually inverse. They are the same tensor represented in different coordinate systems z and z^*, giving the same local distance, where $(\cdot)^*$ means the dual structure of $(\cdot)$.*

Theorem 6.1.2. *The two divergences D and D^* are mutually reciprocal, in the sense of*

$$D^*[y^*:z^*] = D[z:y]. \tag{6.6}$$

The divergence between two points z and y is written in the dual form

$$D[z:y] = k(z) + k^*(y^*) - z \cdot y^*. \tag{6.7}$$

Pythagorean Theorem and Additivity

It has been shown that Space S equipped with a Bergman divergence is Riemannian, but has two dually flat affine structures. This gives rise to the following generalized Pythagorean theorem.

Theorem 6.1.3. ***Pythagorean Theorem:*** *Let P, Q, R be three points in S whose coordinates (and dual coordinates) are represented by z_P,z_Q, z_R (z^*_P, z^*_Q, z^*_R), respectively. When the dual geodesic connecting P and Q is orthogonal at Q to the geodesic connecting Q and R, then*

$$D[P : R] = D[P : Q] + D[Q : R]. \tag{6.8}$$

Dually, when the geodesic connecting P and Q is orthogonal at Q to the dual geodesic connecting Q and R, we have

$$D[R : P] = D[Q : P] + D[R : Q]. \tag{6.9}$$

Proof. By using [6.7], we have

$$\begin{aligned} D[R : Q] + D[Q : P] = & \\ &= k(z_R) + k^*(z^*_Q) + k(z_Q) + K^*(z^*_P) - z_R \cdot z^*_Q - z_Q \cdot z^*_P \\ &= k(z_R) = k^*(z^*_P) + z_Q \cdot z^*_Q - z_R \cdot z^*_Q - z_Q \cdot z^*_P \\ &= D[z_R : z^*_P] + (z_Q - z_R) \cdot (z^*_Q - z^*_P) \end{aligned}$$

The tangent vector of the geodesic connecting Q and R is $z_Q - z_R$, and the tangent vector of the dual geodesic connecting Q and P is $z^*_Q - z^*_P$ in the dual coordinate system. Hence, the second term of the right-hand side of the above equation vanishes, because the primal and dual geodesics connecting Q and R, and Q and P are orthogonal. □

The Pythagorean Theorem cited here is quite important for our goals. It can be seen that it implies additivity of the corresponding divergence functions for two statistically independent subsystems. We view the latter property as necessary for the proper quantitative measure of the degree of belief change – the one we are interested in finding.

6.1.2 Invariant Divergence

Consider again the space S_n of all probability distributions over $n+1$ atoms $X = \{x_0, x_1, ..., x_n\}$. The probability distribution are given by $p = (p_1, p_1, ..., p_n)$, $p_i = \text{Prob}\{x = x_i\}$, $i = 0, 1, ..., n$, $\sum p_i = 1$. We try to define a new divergence measure $D[p : q]$ between two distributions p and q. To this end, the concept of information monotonicity should be introduced.

If we divide X into m groups, $G_1,\ G_2, ...,\ G_m$ $(m < n+1)$, and make sure$X = \cup G_i,\ G_i \cap G_j = \emptyset$. Assume that we do not know the outcomes x_i directly, but can observe which group G_j it belongs to. This is called coarse-graining of X. The coarse-gaining generates a new probability distributions $\bar{p} = (\bar{p}_1, ..., \bar{p}_m)$ over$G_1, ..., G_m$. Let $\bar{D}[\bar{p} : \bar{q}]$ be an induced divergence between $\bar{p}$ and $\bar{q}$. Since coarse-graining summarized some of elements into one group, detailed information of the outcome in each groups is lost. Therefore, it is natural to require

$$\bar{D}[\bar{p} : \bar{q}] \leq D[p : q]. \tag{6.10}$$

For two distributions p and q, assume that the outcome x_i is known to belong to G_i. Then we require more information to distinguish the two probability distributions p and q by knowing further detail inside group G_j. Since x_i belongs to group G_j, we consider the conditional probability distributions

$$p(x_i|x_i \in G_j), \quad q(x_i|x_i \in G_j)$$

inside group G_j. If they are equal, we cannot obtain further information to distinguish p from q by observing elements inside G_j. Hence,

$$\bar{D}[\bar{p} : \bar{q}] = D[p : q]. \tag{6.11}$$

holds, when and only when

$$p(x_i|x_i \in G_j) = q(x_i|x_i \in G_j) \tag{6.12}$$

for all G_j and all $x_i \in G_j$, or

$$\frac{p_i}{q_i} = \lambda_j \tag{6.13}$$

for all $x_i \in G_j$ for some constant λ_j.

All divergence satisfying the above requirements is called an **invariant divergence**, and such a property is termed as information monotonicity.

The f-divergence was introduced by Csiszár [19]. It is defined by

$$D_f[p : q] = \sum p_i f\left(\frac{q_i}{p_i}\right), \tag{6.14}$$

where f is a convex function satisfying $f(1) = 0$. Csiszár found that an $f-$ divergence satisfies information monotonicity. Moreover, the class of f-divergence is unique in the sense that any decomposable divergence satisfying information monotonicity is an f- divergence.

For our goals, the sought for measure of the degree of change in beliefs has to be consistent with the general structure of classical probabilities, i.e. it has to properly respect the conditional structure of the latter. In other words, it the desired measure is a divergence it has to be invariant in the sense of information geometry.

6.1.3 Kullback-Leibler Divergence in Information Geometry View

Consider the set S_n of all discrete probability distributions over $n+1$ elements $X = \{x_0, x_1, ..., x_n\}$. A probability distribution is given by

$$p(x) = \sum_{i=0}^{n} p_i \delta_i(x), \tag{6.15}$$

where $p_i = \text{Prob}\{x = x_i\}$ and $\delta_i(x) = 1, \text{if } x = x_i$ and 0 otherwise. Obviously, $\sum\limits_{i=0}^{n} p_i = 1$. We can use a coordinate system $z = (p_1, p_2, ..., p_n)$ for the set S_n of all such distributions, where $z_0 = p_0$ is regarded as a function of the other coordinates,

$$p_0 = 1 - \sum_{i=1}^{n} z_i. \tag{6.16}$$

The Shannon entropy,

$$H_(z) = -\sum z_i \log z_i - (1 - \sum z_i) \log(1 - \sum z_i), \tag{6.17}$$

is concave, so that $k(z) = -H(z)$ is a concave function of z.

The Riemannian metric induced from $k(z)$ in calculated as

$$g_{ij}(z) = \frac{1}{p_i}\delta_(ij) + \frac{1}{p_0}, \tag{6.18}$$

which is the Fisher information matrix. The divergence function is given by

$$D[z : y] = \sum_{i=0}^{n} z_i \log \frac{z_i}{y_i}. \tag{6.19}$$

which is known as the Kullback-Leibler divergence. It is written in general as

$$D_{KL}[p(x) : q(x)] = \sum_x p(x) \log \frac{p(x)}{q(x)}. \tag{6.20}$$

In [4], it is shown that Kullback-Leibler divergence belongs to the class of Bregman divergences for which the Pythagorean theorem holds. Thus it satisfies the requirement of additivity of our proposed approach.

And if we recall the form of the f-divergence which has the invariance property in equation [6.14] that

$$D_f[p : q] = \sum p_i f(\frac{q_i}{p_i}),$$

we can see that Kullback-Leibler divergence belongs to the f-divergences with $f(u) = u - 1 - \log u$.

Moreover it can be shown [4] that Kullback-Leibler divergence is the only one that is both invariant and dually flat (and thus Bregman and satisfies the Pythagorean theorem) the at the same time on the manifold of probability distributions.

Chapter 7

Conclusion and Future Work

7.1 Final piece of evidence

Let us summarize our findings concerning the main goal of the present thesis – the determination of the proper quantitative measure of the degree of change from the original agent's belief to the updated one. The review of existing literature seems to point rather strongly at Kullback-Leibler divergence as the unique proper quantitative characterization of that degree. Let's give some additional supporting arguments to KL-divergence. Let, as before, the probability measure P describe the agent's original belief which is updated to the measure P^k upon reception of k as the value of the sources answer to the corresponding agent's question. Then the standard total probability rule

$$\sum_k v_k P^k = P$$

has an important interpretation in the given context. It expresses the assumption of rationality of both beliefs – the original and the updated one. In other words, it expresses the observation that the original belief is not erased but simply refined by the new information contained in the answer.

Let

$$L(P) = \int_\Omega \left(f(x_P^*, \omega) - f(x_\omega^*, \omega)\right) \mathrm{d}P(\omega)$$

be the original value of the loss. We are inerested in the sign of change of that loss upon reception of the information contained in the answer $A(Q)$. We can write

$$\begin{aligned}
&L(P) - \sum_k v_k L(P^k) \\
&\qquad = \int_\Omega \left(f(x^*_P, \omega) - f(x^*_\omega, \omega)\right) \mathrm{d}P(\omega) - \sum_k v_k \int_\Omega \big(f(x^*_{P^k}, \omega) - f(x^*_\omega, \omega)\big) \mathrm{d}P^k(\omega) \\
&\qquad = \int_\Omega f(x^*_\omega, \omega) \left(\sum_k v_k \mathrm{d}P^k(\omega) - \mathrm{d}P(\omega)\right) \leftarrow 0 \\
&\qquad + \int_\Omega f(x^*_P, \omega) \mathrm{d}P(\omega) - \sum_k v_k \int_\Omega f(x^*_{P^k}, \omega) \mathrm{d}P^k(\omega) \\
&\qquad = \int_\Omega f(x^*_P, \omega) \left[\sum_k v_k \mathrm{d}P^k(\omega)\right] - \sum_k v_k \int_\Omega f(x^*_{P_k}, \omega) \mathrm{d}P^k(\omega) \\
&\qquad = \sum_k v_k \int_\Omega \left[f(x^*_P, \omega) - f(x^*_{P^k}, \omega)\right] dP^k(\omega) \geq 0
\end{aligned}$$

This result can be formulated in plain language as “Any rational belief update decreases the loss”. Note that here we have not used any quantitative measure of degree of belief change but only the assumption of belief rationality.

Now let us consider the degree of change of belief. If we use KL-divergence as the basis for such a quantity we obtain the following

$$D(P_A \| P) \equiv \sum_k v_k D(P^k \| P) = \sum_k v_k \sum_i P^k(\omega_i) \log \frac{P^k(\omega_i)}{P(\omega_i)}. \tag{7.1}$$

With a little more work it turns out to be possible to show that any higher $D(P_A \| P)$ obtained by question refinement implies a lower loss $L(P_A)$. The main property of KL divergence that makes this claim true is its strong additivity in the sense introduced in earlier chapters.

This establishes a direct link between the quantity and belief change and the loss reduction. Since KL-divergence is the only known measure of difference of two distributions possessing the strong additivity property we obtain an additional hint that it may be the only correct quantity of the degree of change we are looking for.

7.2 Future work

Summarizing the main result of the present thesis we can say that, upon a review of existing measures of information we found the only one possessing all characteristics of the desired measure of the degree of change in the agent's beliefs – the Kullback-Liebler divergence. The next logical step is to use these findings in the development of the quantitative framework of information acquisition optimization. Here we sketch the next few steps that seem rather clear at the time of writing.

- It seems logical, as we have mentioned already, to define the answer depth as the expected value of KL-divergence between the original and the updated belief. This also gives rise to the notion of the question depth defined as the value of the answer depth for the case of a perfect answer.

- The question difficulty functional can then be logically defined as we have already described in the Introduction. The coefficients u in it reflecting the source's knowledge structure are related to the conditional probabilities describing the source's knowledge.

- Then one should explore the possible large scale structures of the source's knowledge in its own right paying special attention to the possible symmetries and the relation of these symmetries to the parameters u of the question difficulty functional.

- Finally, it would be needed to explore the relations between difficulty of questions and the resulting loss especially for cases when the source is not capable of a perfect answer. Does there exist a general (even if approximate) relationship between the difficulty coefficients u and the increase of loss relative to the case of $u = 0$?

Bibliography

[1] J. Aczél and Z. Darózy. *On measures of information and their characterizations.* Academic Press, New York, 1975.

[2] S. I. Amari. Information geometry and its applications: Convex function and dually flat manifold. *Emerging Trends in Visual Computing*, 2009.

[3] S. I. Amari. Differential geometry derived from divergence functions: information geometry approach. *Mathematics of Distances and Applications*, pages 9–23, 2012.

[4] S. I. Amari and A. Cichocki. Information geometry of divergence functions. *Bulletin of the Polish Academy of Sciences: Technical Sciences*, 58(1):183–195, 2010.

[5] A. Ben-Tal, S. Boyd, and A. Nemirovski. Extending scope of robust optimization: Comprehensive robust counterparts of uncertain problems. *Math. Programming, Ser.B*, 107:63–89, 2006.

[6] A. Ben-Tal and A. Nemirovski. Robust convex optimization. *Math. Oper. Res.*, 23(4):769–805, 1998.

[7] J. R. Birge and F. Louveaux. *Introduction to Stochastic Programming.* Springer-Verlag, New York, NY, 1997.

[8] R. F. Bordley. Combining the opinions of experts who partition events differently. *Decision Anal.*, 6(1):38–46, March 2009.

[9] R. F. Bordley. Using bayes' rule to update an event's probabilities based on the outcomes of partially similar events. *Decision Anal.*, 8(2):117–127, June 2011.

[10] L. Bregman. The relaxation method of finding a common point of convex sets and its application to the solution of problems in convex programming. *Comp. Math. Phys.*, USSR7:200–217, 1967.

[11] A. B. Cambel. *Applied Chaos Theory.* Academic Press, London, 1993.

[12] A. Caticha. *Entropic Inference and the Foundations of Physics (monograph commissioned by the 11th Brazilian Meeting on Bayesian Statistics–EBEB-2012.* Sao Paulo: USP Press, 2012.

[13] M. Chaẃez, J. Martinerie, and M. LeVanQuyen. Statistical assessment of nonlinear causality: Application to epileptic eeg signals. *J. of Neurosci. Methods*, 124(2113-128), 2003.

[14] R. Clemen. Combining overlapping information. *Management Sci.*, 33(3):373–380, 1987.

[15] R. Clemen and R. Winkler. Combining probability distributions from experts in risk analysis. *Risk Anal.*, 19(2):187–203, 1999.

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

[17] R. T. Cox. Probability, frequency, and reasonable expectation. *Am. J. Phys*, 14:1–13, 1946.

[18] R. T. Cox. *The Algebra of Probable Inference.* Johns Hopkins Press, Baltimore, 1961.

[19] I. Csiszár. Information-type measures of difference of probability distributions and indirect observations. *Studia Sci. Math.*, 2:299–318, 1967.

[20] D. K. Faddeev. On the concept of entropy of a finite probabilistic scheme. *Uspekhi Mat. Nauk*, 11(1(67)):227–231, 1956.

[21] C. Fox and R. Clemen. Subjective probability assessment in decision analysis: Partition dependence and bias toward the ignorance prior. *Management Sci.*, 51(9):1417–1432, 2005.

[22] S. French. Group consensus probability distributions: A critical survey. *Bayesian Statist*, 2:183–202, 1985.

[23] C. Genest and J. V. Zidek. Combining probability distributions: A critique and an annotated bibliography. *Statist. Sci.*, 1:114–148, 1986.

[24] D. Harmanec and G. J. Klir. Measuring total uncertainty in dempster-shafer theory: A novel approach. *Int. J. Gen. Syst*, 22(4):405–419, 1994.

[25] J. H. Havrda and F. Charvat. Quantification methods of classification processes: Concepts of structural entropy. *Kybernetika*, 3:30–35, 1967.

[26] R. A. Howard. Information value theory. *IEEE Transactions on Systems Science & Cybernetics*, 2(1):22–26, 1966.

[27] S. Jaroszewicz and D. A. Simovici. On axiomatization of conditional entropy of functions between finite sets. *In Proc. 29th ISMVL, Freiburg, Germany*, pages 24–31, 1999.

[28] E. T. Jaynes. Information theory and statistical mechanics i. *Phys. Rev.*, 106:620–630, 1957.

[29] E. T. Jaynes. Information theory and statistical mechanics ii. *Phys. Rev.*, 108:171–190, 1957.

[30] J. N. Kapur. *Measures of Information and Their Applications.* New Age International Limited, Publishers, 1994.

[31] T. Katura, N. Tanaka, A. Obata, H. Sato, and A. Maki. Quantitative evaluation of interrelations between spontaneous low-frequency oscillations in cerebral hemodynamics and systemic cardiovascular dynamics. *NeuroImage*, 31(4):1592–1600, July 2006.

[32] G. J. Klir and D. Harmanec. Generalized information theory: Recent developments. *Kybernetes*, 25(7/8):50–66, 1966.

[33] K. H. Knuth. Lattice duality: The origin of probability and entropy. *Neurocomputing*, 67:245–274, 2005.

[34] S. Kullback. *Topics in Statistical Information Theory.* Springer-Verlag, Berlin, 1987.

[35] Y. Maeda and H. Ichihashi. An uncertainty measure with monotonicity under the random set inclusion. *Int. J. Gen. Syst.*, 21(4):379–392, 1993.

[36] I. I. Mokhov and D. A. Smirnov. El nin o-southern oscillation drives north atlantic oscillation as revealed with nonlinear techniques from climatic indices. *Geophys. Res. Lett.*, 33, 2006. L03708.

[37] E. Perevalov and D. Grace. Towards the full information chain theory: answer depth and source models. *Physical Review E, arXiv:1212.2696v2 [physics.data-an]*, 2013.

[38] E. Perevalov and D. Grace. Towards the full information chain theory: expected loss and information relevance. *Physical Review E, arXiv:1301.2020 [physics.data-an]*, 2013.

[39] E. Perevalov and D. Grace. Towards the full information chain theory: question difficulty. *Physical Review E, arXiv:1212.2693v2 [physics.data-an]*, 2013.

[40] E. Perevalov and D. Grace. Towards the full information chain theory: solution methods for optimal information acquisition problem. *Physical Review E, arXiv:1302.0070 [physics.data-an]*, 2013.

[41] J. R. Pierce. *An Introduction to Information Theory.* Dover Publications, New York, 1980.

[42] J. B. Predd, D. N. Osherson, S. R. Kulkarni, and H. V. Poor. Aggregating probabilistic forecasts from incoherent and abstaining experts. *Decision Anal.*, 5(4):177–189, December 2008.

[43] A. Rényi. *On measures of entropy and information.* Univ. California Press, Berkeley, Calif., 1961.

[44] A. Rényi. *Selected Papers (I, II, III).* Akad. Kiado, Budapest, 1976.

[45] G. Shafer. *A Mathematical Theory of Evidenc.* Princeton University Press, Princeton, NJ, 1976.

[46] C. E. Shannon. A mathematical theory of communication. *Bell Systems Technical Journal*, 27:379–423, 623–659, 1948.

[47] D. A. Simovici and S. Jaroszewicz. An axiomatization of partition entropy. *IEEE Trans. Inf. Theory*, 48(7):2138–2142, July 2002.

[48] I. N. Taneja. On generalized information measures and their applications. *Advances in Electronics and Electron Physics*, 76:327–413, 1989.

[49] P. F. Verdes. Assessing causality from multivariate time series. *Phys.Rev.E, 72*, 2005. 02622.

[50] P. A. Viola. Alignment by maximization of mutual information. A.i. technical report 1548, Massachusetts Institute of Technology, June 1995.

Biography

Zhenglin Wei was born on December 27^{th}, 1992, in Zibo City, Shandong Provence, P. R. China. He obtained his bachelor of science degree in Systems Science and Engineering(Financial Engineering) at University of Shanghai for Science and Technology, P. R. China. He has been a graduate student in the Department of Industrial and Systems Engineering at Lehigh University since 2015 and expects to complete M.S. degree in Industrial and Systems Engineering in 2017.