Information-Theoretic Analysis using Theorem Proving

Tarek Mhamdi

A Thesis

in

The Department

of

Electrical and Computer Engineering

Presented in Partial Fulfillment of the Requirements

for the Degree of Doctor of Philosophy at

Concordia University

Montréal, Québec, Canada

December 2012

CONCORDIA UNIVERSITY

Division of Graduate Studies

This is to certify that the thesis prepared

By: **Tarek Mhamdi**

Entitled: **Information-Theoretic Analysis using Theorem Proving**

and submitted in partial fulfilment of the requirements for the degree of

Doctor of Philosophy

complies with the regulations of this University and meets the accepted standards with respect to originality and quality.

Signed by the final examining committee:

____________________ Dr. Deborah Dysart-Gale

____________________ Dr. Amy Felty

____________________ Dr. Nizar Bouguila

____________________ Dr. Weiping Zhu

____________________ Dr. Otmane Ait Mohamed

____________________ Dr. Sofiène Tahar

Approved by ______________________________

Chair of the ECE Department

________ 2012 ______________________________

Dean of Engineering

ABSTRACT

Information-Theoretic Analysis using Theorem Proving

Tarek Mhamdi

Concordia University, 2012

Information theory is widely used for analyzing a wide range of scientific and engineering problems, including cryptography, neurobiology, quantum computing, plagiarism detection and other forms of data analysis. Despite the safety-critical nature of some of these applications, most of the information-theoretic analysis is done using informal techniques, mainly computer simulation and paper-and-pencil analysis, and thus cannot be completely relied upon. The unreliable nature of the produced results poses a serious problem in safety-critical applications and may result in heavy financial losses or even the loss of human life. In order to overcome the inaccuracy limitations of these techniques, this thesis proposes to conduct the analysis within the trusted kernel of a higher-order-logic (HOL) theorem prover. For this purpose, we provide HOL formalizations of the fundamental theories of measure, Lebesgue integration and probability and use them to formalize some of the most widely used information-theoretic principles. We use the Kullback-Leibler divergence as a unified measure of information which is in turn used to define the main measures of information like the Shannon entropy, mutual information and conditional mutual information. Furthermore, we introduce two new measures of information leakage, namely the information leakage degree and the conditional information leakage degree and compare them with existing measures. We illustrate the usefulness of the proposed framework by tackling various applications including the performance analysis of a communication encoder

used in the proof of the Shannon source coding theorem, the quantitative analysis of privacy properties of a digital communications mixer and the one-time pad encryption system using information-theoretic measures.

To My Wife and Daughter, My Mom and Dad, My Sisters and Brother.

ACKNOWLEDGEMENTS

First, I am deeply grateful to Dr. Sofiène Tahar for his help, guidance and encouragement throughout my graduate studies. I could not have wished for a better thesis supervisor. Many thanks to Dr. Osman Hasan for his support in my research and for his encouragement. I would like to thank the love of my life, Souha Mahmoudi, for her unconditional love and support. I am deeply grateful to my parents for all the love and support they have provided me over the years. Nothing would be possible without them. Many thanks to the members of the thesis committee for the assistance they provided at all levels of the research project. Finally, many thanks to my good friends at the hardware verification group for their support and motivation.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ACRONYMS

ACL2	A Computational Logic for applicative common Lisp
AEP	Asymptotic Equipartition Property
CDF	Cumulative Distribution Function
DPI	Data Processing Inequality
HOL	Higher-Order Logic
HOL4	HOL4 Theorem Prover
LCF	Logic of Computable Function
KL	Kullback-Leibler
LISP	LISt Processing
ML	Meta Language
OTP	One-Time Pad
PDF	Probability Density Function
PMF	Probability Mass Function
PRISM	PRobabilistIc Symbolic Model checker
PVS	Prototype Verification System
TOR	The Onion Router
WLLN	Weak Law of Large Numbers
ZFC	Zermelo-Fraenkel with axiom of Choice

Chapter 1

Introduction

1.1 Motivation

Knowing ignorance is strength. Ignoring knowledge is sickness.

Lao-Tse, Tao Te Ching [6th Century BCE]

Hardware and software systems usually exhibit some kind of random or unpredictable behavior, either because of a faulty component, interaction with an unpredictable environment or simply because of the use of a probabilistic algorithm within the system. Due to these elements of randomness and uncertainty, establishing the correctness of a system under all circumstances usually becomes impractically expensive. The engineering approach to evaluate the performance of these systems is to use probabilistic analysis. Furthermore, information-theoretic analysis is gaining more ground in the study of correctness and performance of a broad range of scientific and engineering systems. In fact, after it was first introduced by Shannon in his seminal paper [62], information theory has become an increasingly popular discipline in a wide range of applications including communication, cryptography, quantum computing, plagiarism detection and other forms of data analysis.

Information-theoretic analysis consists in using measures of information to quantify the flow of information and evaluate the performance of a system or a protocol. Examples of this analysis include the evaluation of anonymity networks and security protocols. Anonymity networks such as Crowds [54] and Tor [20] have been proposed to provide anonymous communication between entities in a network. Analyzing the anonymity properties of these protocols consists in finding out how much information an attacker can learn about the senders and receivers in the network. One way to do so is through quantitative analysis of information flow [64, 58] which allows to measure how much information about the high security inputs of a system can be leaked, accidentally or maliciously, by observing the systems outputs and possibly the low security inputs. Quantitative analysis of information flow has also been proposed to analyze security protocols [64]. In fact, while these protocols aim to preserve sensitive and confidential data and prevent it from being leaked or tainted, a small leakage of information is sometimes necessary, as is the case for password checking or for a voting protocol.

Traditionally, computer simulations and paper-and-pencil based analysis have been used for the analysis of probabilistic systems as well as the quantitative analysis of information. Computer simulation provides, however, less accurate results due to the usage of computer arithmetics, such as floating or fixed point numbers, that leads to numerical approximations. In addition, it cannot handle large-scale problems due to their enormous computer processing time requirements. The unreliable nature of the produced results poses a serious problem in safety-critical applications, such as those in secure communications, space travel, military applications, and medicine. Paper-and-pencil analysis, on the other hand, does not scale well to complex systems and is prone to human error.

As an alternative approach to overcome the above shortcomings, we propose a computer-assisted technique which consists in conducting the probabilistic and information-theoretic analysis using formal methods [2]. Formal methods are techniques used to model complex software and hardware systems as mathematical entities. They are used for the specification, analysis and verification of these systems to improve their reliability, design time and comprehensibility. They broadly fall into two main categories: *proof based* methods, mainly theorem proving and *state-exploration* methods, mainly model checking. While theorem proving is a scalable technique that can handle large designs, model checking suffers from the so called state-explosion problem which prevents its application to larger systems [40]. On the other hand, while model checking is fully automatic, deriving proofs is a user guided technique that requires a lot of expertise and hence can be tedious and difficult.

Several probabilistic model checking tools have been developed, we mention for example, PRISM [41] and VESTA [60]. However, in addition to the above shortcomings of model checking, this technique can only be used for systems that can be expressed as finite state machines or Markov chains. Higher-order-logic, on the other hand, is highly expressive and can be used to describe any mathematical relationship, in particular, the mathematical theories needed to conduct the probabilistic and information-theoretic analysis. For this reason, we propose to conduct this analysis within the sound core of a higher-order-logic theorem prover

Theorem proving [30] is a field of computer science and mathematical logic that allows to conduct computer-assisted formal proofs of the correctness of systems and programs using mathematical reasoning. The implementation and specification of a system are both expressed in terms of logical formulas and the proof of correctness is derived from a very small set of axioms and inference rules. The soundness of theorem proving guarantees that only valid results are provable, hence, overcoming the inaccuracies of simulation and paper-and-pencil based techniques. We give a brief overview of theorem proving, which we use in our work, in Section 2.2.

In order to achieve our objective of using a higher-order-logic theorem prover to perform the information-theoretic analysis, we need first to formalize, or write in a formal language, all the underlying theories that are needed to express the systems and protocols under consideration. This includes the formalization of probability and information theory concepts in higher-order logic. In this work, we propose a generalized higher-order-logic formalization of the underlying mathematical theories of measure [9], Lebesgue integration [6], probability [26] and information theory [14]. Using measure theory to formalize probability has the advantage of providing a mathematically rigorous treatment of probability and a unified framework for discrete and continuous probability measures. Lebesgue integration is used to develop statistical properties of random variables and various measures of information.

Several measures of information flow have been proposed in the literature. For instance, Serjantov [61] and Diaz et al. [19] independently proposed to use the entropy to define the quality of anonymity and to compare different anonymity systems. Malacaria [44] defined the leakage of confidential information in a program as the conditional mutual information between its outputs and secret inputs, given the knowledge of its low security inputs. Deng [18] proposed relative entropy as a measure of the amount of information revealed to the attacker after observing the outcomes of the protocol, together with the a priori information. Chatzikokolakis [10] modeled anonymity protocols as noisy channels and used the channel capacity as a measure of the loss of anonymity. Zhu and Bettati [68] used the mutual information to define what they called *anonymity degree* and used it to analyze a digital MIX, which is a communication system introduced by Chaum [11] to create hard-to-trace communications. We introduce two novel measures of information leakage, namely the information leakage degree and the conditional information leakage degree. We will compare them to the existing measures and show that they have the advantage that they not only quantify the information leakage but also describe the quality of leakage by normalizing the measure by the maximum leakage that the system allows under extreme situations. We show how the information leakage degrees can be used to evaluate both the anonymity and privacy properties of protocols. We compare the proposed information leakage degree to the anonymity degree introduced in [68] and show that our definition is more generic.

We illustrate the practical effectiveness of our work and its utilization to conduct information-theoretic analysis using a theorem prover, by tackling various applications including a data compression [14] application consisting of the proof of the Shannon source coding theorem. We also evaluate the anonymity properties of an anonymity-based MIX channel [68] as well as the privacy properties of the one-time pad encryption system [15].

In this thesis, we use the HOL4 theorem prover [24] for the above mentioned formalization and verification tasks. The main motivation behind this choice is to build upon existing formalizations of measure [37] and Lebesgue integration [13] theories in HOL. The methodology is, however, valid for other higher-order-logic theorem provers.

1.2 State-of-the-Art

Probabilistic analysis using formal methods have been an active research area, lately. The most mature technique has been probabilistic model checking where several tools have been developed. The formalization of probability and some concepts of information theory in proof assistants have also been investigated in several related works. Below, we present the state-of-the-art in terms of the different theories that have been formalized as well as a brief overview of probabilistic model checking.

1.2.1 Measure and Probability

The early foundations of probabilistic analysis in a higher-order-logic theorem prover were laid down by Nędzusiak [46] and Bialas [8] when they proposed a formalization of some measure and probability theories in higher-order logic. Hurd [37] implemented their work and developed a formalization of measure and probability theories in HOL. Despite important contributions, Hurd's formalization did not include basic concepts such as the expectation of random variables. In Hurd's formalization, a measure space is the pair $(\mathcal{A}, \mu)$; $\mathcal{A}$ is a set of subsets of a space X, called the set of measurable sets and μ is a measure function. Hence, the space is implicitly the universal set of the appropriate type. This approach does not allow to construct a measure space where the space is not the universal set. The only way to apply this approach for an arbitrary space X is to define a new type for the elements of X, redefine operations on this set and prove properties of these operations. This requires considerable effort that needs to be done for every space of interest.

Hasan [31] built upon Hurd formalizations of measure and probability to verify the probabilistic and statistical properties of some commonly used discrete [33] and continuous [32] random variables. The results were then utilized to formally reason about the correctness of many real-world systems including the analysis of the Coupon Collector's problem [34] and the Stop-and-Wait protocol [35]. Hasan's work inherits the above mentioned limitations of Hurd's work. For example, separate frameworks for handling systems with discrete and continuous random variables are required and the inability to handle multiple continuous random variables. Another important limitation of this work is the requirement of independence of random variables. This assumption cannot be satisfied for a large class of systems involving multiple random variables.

Abbasi [1] extended the work of Hasan by formalizing statistical properties of continuous random variables as well as the probability distribution properties of multiple random variables and used it for the formal reliability analysis of engineering systems using theorem proving. The results from Abbasi's formalization are valid only for the specific probability distributions considered. In fact, to be able to prove a property of a specific random variable, the user would start with that random variable and derive the proof. A better approach would be to prove a general result that can be applied to various random variables. Furthermore, Abbasi's work has also the disadvantage that the results cannot be used for both discrete and continuous random variables and inherits the limitation of requiring the independence of random variables from Hasan's formalization.

Based on the work of Hurd [37], Coble [13] extended the formalization of measure theory by defining the measure space as the triple $(X, \mathcal{A}, \mu)$ allowing him to work with an arbitrary space X and hence eliminate the above shortcoming of the formalization of Hurd. Coble has also defined probability spaces and random variables. However, this formalization considers only finitely-valued measures and functions as it was based on standard real numbers. Using extended-real numbers in the formalization has many advantages. It allows us to define sigma-finite and other infinite measures as well as signed measures. It also allows to define extended-real valued functions. Furthermore, Coble's formalization does not include Borel spaces and hence it was not possible to prove the properties of measurable functions. A later version of the formalization of measure contained the definition of the Borel sigma algebra but it was defined based on open intervals instead of open sets. This limits the applications to real-valued measurable functions.

1.2.2 Lebesgue Integration

Richter [55] ported Hurd's formalization [37] of measure theory and probability theory in Isabelle/HOL [51], and used it to formalize Lebesgue integration. Only real-valued functions are considered in this work as the definition of Borel spaces was also defined as being generated by the intervals. Defining the Borel sigma algebra based on the open sets allows to work with functions defined on any topological space, such as the complex numbers or $\mathbb{R}^n$, the n-dimensional Euclidean space. Richter's formalization does include the main convergence theorems of the Lebesgue integral and the important Radon Nikodym derivative, which used to define several concepts of probability and information theories. This formalization does not support extended-valued real functions.

Coble has also provided a nice formalization of Lebesgue integration in the HOL4 theorem prover but it was not based on the extended-real numbers and hence limiting the scope of applications and, more importantly, it prevents us from proving various convergence theorems and the important Radon Nikodym theorem. Furthermore, in the formalization of Coble which lacked the Borel sigma algebra, it was not possible to prove properties of the Lebesgue integral for arbitrary functions like the monotonicity and linearity of the integral. Only the properties for positive simple functions were provided. Finally, defining the Borel sigma algebra using open sets, allows to prove the properties and apply the Lebesgue integral to a large class of continuous functions, in particular, trigonometric and exponential functions.

A formalization of the Lebesgue integral on the extended-reals has been proposed in Mizar [63]. To the best of our knowledge, the Radon Nikodym derivative and its properties as well as the Lebesgue convergence theorems have not been formalized in Mizar. Finally, in his work on the formalization of topology using the PVS theorem prover, Lester [42] provided formalizations for measure and integration theories. Lester's formalization lacks the proofs of the properties of the Lebesgue integral as well as the Lebesgue convergence theorems, both of which are very important to the usability of the formalization to analyze systems properties.

1.2.3 Information Theory

Coble used his formalization of measure theory and probability to formalize some concepts of information theory in HOL. This formalization, evidently, inherits the above mentioned drawbacks of his formalization measure and Lebesgue integration. In fact, the lack of extended-real numbers in his formalization, as mentioned above, prevented him from proving the important Radon Nikodym theorem. This theorem plays a vital role in the proof of existence of the Radon Nikodym derivative and hence allows to prove various properties of the derivative and, by extension, all the measures of information that use it in their definitions.

A formalization of the positive extended-reals in HOL was proposed by Hurd [38] and has been imported to the Isabelle theorem prover. Hölzl used this work to formalize various measures of information and the underlying theories in Isabelle, based on the work of Coble. While this is a work in progress, the formalization that has been published by Hölzl [36] lacks the important properties of the Radon Nikodym derivative and the Kullback-Leibler divergence and the different measures of information. It also does not handle signed measures or functions taking negative valued as only positive extended-real numbers are supported.

More recently, Affeldt [3] provided a simplified formalization of probability theory in the Coq proof assistant [7] and used it to formalize basic concepts of information theory. The aim of this work was to prove Shannon's source coding and channel theorems and not to provide a generalized formalization of probability or information theory that can be used to analyze other applications.

1.2.4 Probabilistic Model Checking

In addition to theorem proving, probabilistic model checking is the second most widely used formal probabilistic analysis method [4, 57]. Like traditional model checking [5], probabilistic model checking involves the construction of a precise state-based mathematical model of the given probabilistic system, which is then subjected to exhaustive analysis to verify if it satisfies a set of probabilistic properties formally expressed in some appropriate logic. Numerous probabilistic model checking algorithms and methodologies have been proposed in the open literature, e.g., [17, 50], and based on these algorithms, a number of tools have been developed, e.g., PRISM [41] and VESTA [60].

In addition to the accuracy of the results, another important feature of probabilistic model checking is the ability to perform the analysis automatically. On the other hand, probabilistic model checking is limited to systems that can only be expressed as probabilistic finite state machines or Markov chains. Another major limitation of the probabilistic model checking approach is state-space explosion [5]. Similarly, to the best of our knowledge, it has not been possible to precisely reason about information-theoretic foundations, such as expectation, variance and measures of information, using probabilistic model checking so far. Higher-order-logic theorem proving, on the other hand, overcomes the limitations of probabilistic model checking and thus allows conducting formal information-theoretic analysis of a wide range of engineering systems but at the cost of significant user interaction.

1.3 Proposed Methodology

The main objective of this work is to provide a comprehensive framework to conduct probabilistic and information-theoretic analysis of systems and protocols within the sound core of a theorem prover, as an alternative to less accurate techniques like simulation and paper-and-pencil methods and to other less scalable techniques like probabilistic model checking. We provide the tools to model random components of systems and protocols to be able to prove their desired probabilistic, statistical and information theoretic properties. As depicted in Figure 1.1, the proposed framework provides the necessary tools to mathematically express a system description and its

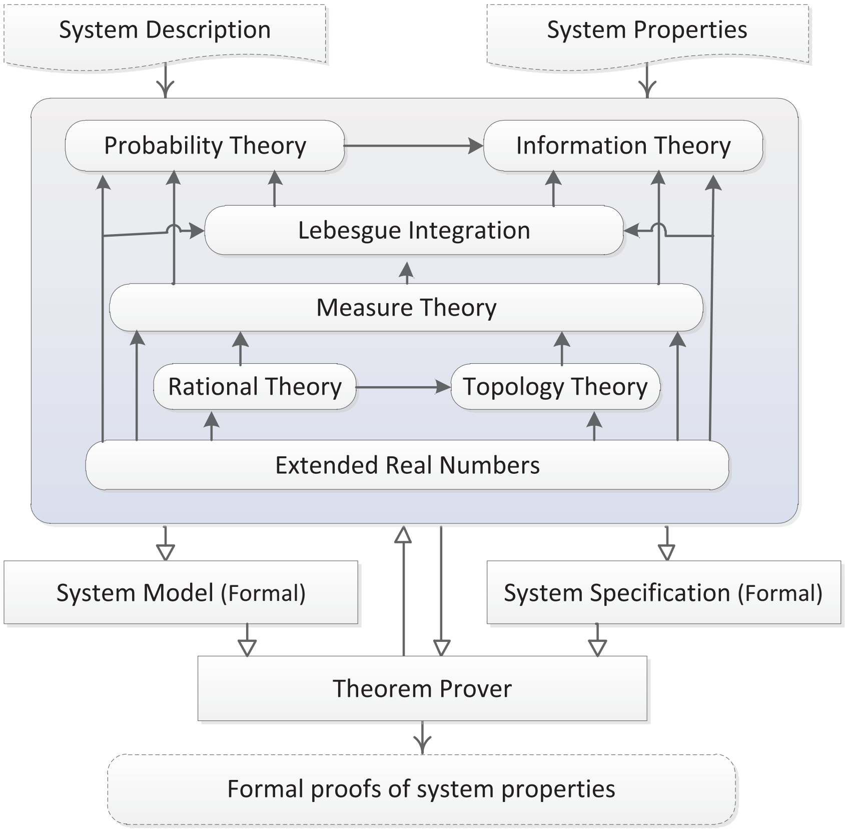

Figure 1.1: Overview of the Proposed Framework

desired properties in a format that is understood by the theorem prover. The framework is presented in terms of theories to be developed specifically for this work but are usable in a much wider range of applications. Lebesgue integration, for instance, is used in this work to define statistical properties of random variables but the developed theory can also be used in the study of Fourier series and Fourier transforms.

We provide a formalization in HOL of the set of extended-real numbers $\overline{\mathbb{R}}$, which is the set of real numbers augmented by the negative and positive infinity. We use this formalization as a basis for the development of the various theories of the framework allowing us to prove several properties, mainly convergence theorems, that would not have been possible to prove using the normal (finite) real numbers. We also provide a formalization of the set of rational numbes $\mathbb{Q}$ as well as some concepts of topology of $\overline{\mathbb{R}}$, necessary in our development of measure theory.

We use measure theory to formalize probability as it has the advantage of providing a mathematically rigorous treatment of probability and a unified framework for discrete and continuous probability measures. We formalize measure theory based on the Borel spaces allowing us to work on any topological space and prove important properties of extended-real-valued measurable functions. To do so, we make use of the formalization of $\mathbb{Q}$ and the topology of $\overline{\mathbb{R}}$ to formalize the Borel sigma algebra in terms of open sets.

We formalize the Lebesgue integral in HOL based on the extended-real numbers. Using the theories mentioned above, we prove various properties of the integral, especially its convergence theorems. The use of extended-real numbers allows us to prove the important Radon Nikodym theorem and use it to define the Radon Nikodym derivative, necessary to define most commonly used measures of information. We use the Lebesgue integral to define various statistical properties of random variables, such as the expectation and variance, and the different measures of information.

We formalize probability in higher-order logic according to the Kolmogorov axiomatic definition of probability [39]. This definition provides a mathematically consistent way for assigning and deducing probabilities of events. It consists in defining a set of all possible outcomes, Ω, called the sample space, a set F of events which are subsets of Ω and a probability measure p such that (Ω, F, p) is a measure space with $p(\Omega) = 1$. In this context, a random variable is then a measurable function and its expectation is equal to the Lebesgue integral with respect to the probability measure.

The main goal of our framework is the information theory. The formalization of information theory consists in using the underlying theories of measure, Lebesgue integration and probability to develop a higher-order-logic formalization of the main concepts and measures of information. We make use of the Radon Nikodym derivative, defined in the Lebesgue integration theory, to formalize the Kullback-Leibler divergence. The latter is then used to define most commonly used measures of information such us the Shannon entropy, mutual information and conditional mutual information.

1.4 Thesis Contributions

The main contribution of this thesis is an approach and the required formal framework for conducting the information-theoretic analysis within the trusted kernel of a higher-order-logic theorem prover. We propose this technique as an alternative approach to less accurate or less scalable techniques like computer simulation and paper-and-pencil analysis. To achieve this goal, we formalize several fundamental mathematical theories including measure theory, Lebesgue integration, probability and information theory. Each of these underlying theories constitute a considerable contribution to the theorem proving community as they can be used in a wide range of engineering and mathematical applications. They have been released in the official distribution of the HOL4 theorem prover [47] but they can also be adapted to any other higher-order-logic theorem prover. We list below the main contributions of this work with references to related publications provided in the Biography section at the end of the thesis.

- Formalization of the extended-real numbers in HOL including the type and operators definition as well as their properties. This formalization is used to define the various theories of this work, but can be useful to develop a number of other mathematical theories [Bio-Cf2].

- Formalization of measure theory over the extended-real numbers which allows us to work with non-negative measures, signed measures as well as sigma-finite and other infinite measures. This theory includes the formalization of Borel spaces based on open sets making it possible to define measurable functions over any topological space. This required the formalization of basic concepts of the topology of the set of real numbers as well as a rich formalization of the rational numbers in HOL [Bio-Cf2, Bio-Cf3].

- Formalization of Lebesgue integration over extended-real-numbers in HOL including the integral definition and its properties for arbitrary functions. The use of extended-real numbers allowed us to prove various properties and convergence theorems that would not have been possible to prove with normal real numbers [Bio-Cf2, Bio-Cf3].

- Formalization of probability theory in higher-order-logic including probability spaces, random variables and probability mass functions. We used the Lebesgue integration to define the statistical properties of random variables such as the expectation, variance and covariance. We proved some classical results from the probability theory including the Markov and Shannon inequalities as well as the Weak Law of Large Numbers [Bio-Jr2]

- Formalization of the most commonly used measures of information including the Shannon entropy, mutual information and conditional mutual information. We use the Kullback-Leibler divergence as a unified measure of information from which we derive all the other measures and provide simpler expressions of these measures for the case of finite spaces. We proved the Asymptotic Equipartition Property, an important result in information theory used in the proof of several theorems such as the Shannon source coding theorem [Bio-Jr2]. To the best of our knowledge, this is the first higher-order-logic formalization of these information-theoretic notions which also includes their properties.

- An approach to conduct quantitative analysis of information flow using a theorem prover and proposed two new measures of information leakage. We used this technique to analyse the anonymity properties of an anonymity-based single

MIX [Bio-Cf1] and later extended it to study the security performance of the one-time pad encryption system [Bio-Jr1].

1.5 Thesis Organization

The rest of the thesis is organized as follows: In Chapter 2, we provide a brief overview of information theory starting from quantifying uncertainty to the definition of the different measures of information. We also provide in this chapter an introduction to theorem proving and the HOL4 theorem prover.

In Chapter 3, we present a formalization of the fundamental theories of measure and Lebesgue integration in HOL, based on the extended-real numbers. In measure theory, we formalize the basic definitions as well as the Borel spaces and use them to verify the properties of measurable functions. Finally, we present our formalization of the Lebesgue integral and prove its main properties.

The formalization of probability concepts is presented in Chapter 4 including the basic definitions of probability spaces and random variables as well as their statistical properties. We also provide a detailed formalization of the most commonly used measures of information and their properties. Finally we provide an overview of the field of quantitative analysis of information flow and propose two novel measures of information leakage.

We use the proposed approach of information-theoretic analysis and the developed formal framework in the study of various applications in Chapter 5. We prove the properties of the typical encoder as well as analyze an anonymity-based single MIX. Finally, we evaluate the performance of the one-time pad encryption.

Finally, Chapter 6 provides concluding remarks and several future research directions.

Chapter 2

Preliminaries

In this chapter, we provide a brief overview of information theory. We start by linking uncertainty to information and use that observation to describe the different measures of information that have been proposed to conduct the information-theoretic analysis. We also provide a short introduction to higher-order-logic theorem proving technology and the HOL4 theorem prover, the proof assistant we used in our development.

2.1 Information Theory

> "*Real knowledge is to know the extent of ones ignorance.*"
>
> Confucius [551-479 BCE]

This statement captures concisely the relationship between knowledge or information and uncertainty. In fact, information has been intuitively defined as the reduction in uncertainty. As a result, to come up with a measure of information, we first need to be able to quantify uncertainty.

2.1.1 Quantifying Uncertainty

Any situation of uncertainty is characterized by a number n of possibilities where it is unknown which one will be selected. Formally, such a situation can be described by a set of possibilities $S = \{e_1 \ldots e_n\}$ called a *scheme of choice*. We are interested in coming up with a measure of the uncertainty in the scheme of choice S. Intuitively, the uncertainty can be measured by the number of questions that are needed to be asked to determine which possibility was selected. A natural choice would be the cardinality $|S|$ or the number of elements of S. A more clever choice, however, is $\log |S|$, which is the number of questions needed if we recursively partition the set into two halves of equal size, if $|S|$ is even. Otherwise, the number of questions needed is $\log |S| + 1$.

When the probabilities of the different outcomes are known, it is possible to come up with a measure that better describes the uncertainty induced by the different outcomes. First, the uncertainty measure should be a decreasing function of the probability. In fact, the more likely the occurrence of a particular event is, the less anticipating uncertainty its actual observation contains. Furthermore, the uncertainty of a particular joint outcome should be equal to the sum of the uncertainties of the individual outcomes when the outcomes are independent. This property is a form of a Cauchy equation and the solution should be from the class of functions defined by the equation $h(x) = c \log_b(p(x))$, where c is an arbitrary constant and b is a non-negative constant distinct from 1. The constant c should be negative in order for the measure to be decreasing with probability. Finally, since the uncertainty is maximized when all the outcomes are equiprobable c can be chosen to be -1. The choice of b determines the unit of uncertainty: *bit* when $b = 2$ and *nat* when $b = 10$.

The uncertainty associated with the probabilistic scheme of choice is the average of the uncertainties of all the possible outcomes, weighted by their corresponding probabilities. This definition coincides with the definition of Shannon Entropy.

$$H(X) = \sum -p(x)log(p(x))$$

2.1.2 Measures of Information

Information is defined as the amount of reduced uncertainty. Consider a random variable X taking values $x \in S$ with probabilities $p(x)$. The random variable describes an experiment which outcome is uncertain. The uncertainty in this case is equal to the entropy $H(X) = -\sum_{x \in S} p(x) \log(p(x))$. When a certain outcome is observed, the uncertainty is eliminated and the information gained by performing the experiment is equal to $H(X)$. The entropy of a random variable is then a measure of the amount of information gained by observing the actual value of the variable.

If the experiment is carried out only partially where the actual value is not observed but instead some event E occurs. The amount of information gained in this case is equal to the initial uncertainty $H(X)$ reduced by the posterior uncertainty $H(X|E)$. This definition can, however, result in negative information. More generally, the reduction in uncertainty of a random variable X due to another random variable Y, called the *mutual information*, is equal to $I(X;Y) = H(X) - H(X|Y)$.

$$I(X;Y) = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$

The mutual information is a symmetric and non-negative measure of the dependence between the two random variables. It is a special case of a more general quantity called the *relative entropy* or *Kullback-Leibler (KL) divergence*. The latter is a measure of the distance between two probability distribution functions p and q and is defined as:

$$D(p||q) = \sum_{x} p(x) \log \frac{p(x)}{q(x)}$$

Using this definition, the mutual information is actually the KL divergence between the joint probability $p(x, y)$ and the product of marginal probabilities $p(x)p(y)$. If the two random variables are independent ($p(x, y) = p(x)p(y)$), the divergence is equal to zero.

Information theory [62, 25] was developed as a mathematical theory for communication by Claude E. Shannon to define the theoretical limits on the achievable performance of data compression and transmission rate of communication. The limits, being the entropy and the channel capacity, respectively, are given in terms of coding theorems for information sources and noisy channels. Information theory has since been used in analyzing the correctness and performance of a broad range of scientific and engineering systems.

A higher-order-logic formalization of the most commonly used measures of information is presented in Chapter 4, based on the definition of the KL divergence. We also prove, in Chapter 5, the Asymptotic Equipartition Property (AEP) and use it in the proof of the Shannon source coding theorem.

2.2 Theorem Proving

Theorem proving is an approach where both the system and its desired properties are expressed as formulae in some mathematical logic. This logic is defined by a formal system, called *proof system* or *calculus*, which defines a set of *axioms* and a set of *inference rules*. Theorem proving is the process of deriving formal proofs from the basic axioms and possibly intermediate lemmas using inference rules. The axioms are usually "elementary" in the sense that they capture the basic properties of the logic's operators.

Proof styles are often characterized as "forward" or "backward". A forward proof starts with the axioms and assumptions; inference rules are then applied until the desired theorem is proven. A backward proof starts with the theorem as a goal and tactics are applied to reduce the goal into simpler intermediate sub-goals. Sufficiently simple sub-goals are discharged by matching axioms or assumptions or by applying built-in decision procedures.

Many theorem-proving systems have been implemented and used for all kinds of verification problems. The most popular proof assistants include HOL4 [24], Isabelle [51], ACL2 [43], PVS [48] and Coq [7]. These systems are distinguished by, among other aspects, the underlying mathematical logic, the way automatic decision procedures are integrated into the system and the user interface. We use the HOL4 theorem prover in our work, mainly because we build on several underlying theories that exist in HOL4.

HOL4 Theorem Prover

The HOL4 system is a general purpose theorem prover which is capable of conducting proofs in higher-order logic. It utilizes the simple type theory of Church [12] along with Hindley-Milner polymorphism [45] to implement higher-order logic. HOL has been successfully used as a verification framework for both software and hardware as well as a platform for the formalization of pure mathematics.

In order to ensure secure theorem proving, the logic in the HOL system is represented in the strongly-typed functional programming language ML [52]. An ML abstract data type is used to represent higher-order-logic theorems and the only way to interact with the theorem prover is by executing ML procedures that operate on values of these data types. The HOL core consists of only 5 basic axioms and 8 primitive inference rules, which are implemented as ML functions. Soundness is assured as every new theorem must be verified by applying these basic axioms and primitive inference rules or any other previously verified theorems/inference rules.

HOL4 has automatic recursive type definitions, structural induction tools and rewriting tools. The set of types, type operators, constants and axioms available in HOL4 are organized in the form of *theories*. There are two built-in primitive theories, *bool* and *ind*, for Booleans and individuals, respectively. Other important theories, which are arranged in a hierarchy, have been added to axiomatize lists, products, sums, numbers, primitive recursion and arithmetic. On top of these, users are allowed to introduce application-dependent theories by adding relevant types, constants, axioms and definitions.

HOL4 supports both forward and goal-directed backward proofs in a natural-deduction style calculus. The user interacts with HOL4 through the functional meta-language ML and proofs are derived by applying tactics to proof goals. A tactic corresponds to a high-level proof step which automatically generates a sequence of elementary inference rules necessary to justify the step.

The HOL theorem prover includes many proof assistants and automatic proof procedures [27] to assist the user in directing the proof. The user interacts with a proof editor and provides it with the necessary tactics to prove goals while some of the proof steps are solved automatically by the proof decision procedures. Table 2.1 provides the mathematical interpretations of some frequently used HOL symbols and functions, which are inherited from existing HOL theories, in this thesis.

Table 2.1: HOL Symbols and Functions

HOL Symbol	**Meaning**
$\wedge$	Logical *and*
$\vee$	Logical *or*
$\neg$	Logical *negation*
::	Adds a new element to a list
++	Joins two lists together
`hd L`	Head element of list L
`tl L`	Tail of list L
$(\mathtt{a},\mathtt{b})$	A pair of two elements
`fst`	First component of a pair
`snd`	Second component of a pair
$\lambda\mathtt{x}.\mathtt{t}(\mathtt{x})$	Function that maps x to $t(x)$
$\{\mathtt{x}\|\mathtt{P}(\mathtt{x})\}$	Set of all x such that $P(x)$
$\{\mathtt{X}\|\mathbb{T}\}$ *Or* `Univ`	Universal Set
$\{\mathtt{X}\|\mathbb{F}\}$ *Or* $\{\}$	Empty Set
`FINITE s`	s is a finite set
`compl A`	Complement of set A
`A subset B`	A is a subset of B
`A inter B`	A intersection B
`A union B`	A union B
`A diff B`	Difference between sets A and B
`e INSERT s`	{e} union s
`s DELETE e`	s diff {e}
`disjoint A B`	Sets A and B are disjoint
`image f A`	Set with elements $f(x)$ for all $x \in A$
`bigunion P`	Union of all sets in the set P
$\mathtt{sum}(0,\mathtt{k})(\lambda\mathtt{n}.\mathtt{f}(\mathtt{n}))$	$\sum_{n=0}^{k-1} f(n)$
$\mathtt{suminf}(\lambda\mathtt{n}.\mathtt{f}(\mathtt{n}))$	$\lim\limits_{k\to\infty} \sum_{n=0}^{k} f(n)$
$\mathtt{lim}(\lambda\mathtt{n}.\mathtt{f}(\mathtt{n}))$	Limit of a *real* sequence f
$@\mathtt{x}.\mathtt{t}(\mathtt{x})$	Some x such that t(x) is true
`CHOICE s`	Some x in s

Chapter 3

Measure Theory and Lebesgue Integration in HOL

In order to provide a formalization of probability and information theory in higher-order logic, we need first to formalize the fundamental theories of measure and Lebesgue integration.

We use measure theory to formalize probability and information theory as it has the advantage of providing a mathematically rigorous treatment of probabilities and a unified framework for discrete and continuous probability measures. In this context, a probability measure is a measure function, an event is a measurable set and a random variable is a measurable function.

We use the Lebesgue integral to define the statistical properties of random variables such as the expectation, variance and different measures of information. The reason behind this choice over the more commonly known Riemann integral is the unified definition of the Lebesgue integral for both discrete and continuous cases as well as the ability to handle a broader class of functions. Furthermore, the Lebesgue integral exhibits a better behavior when it comes to interchanging limits and integrals.

Both formalizations of measure theory and Lebesgue integral that we will present in the next sections are based on the extended-real numbers. This allows us to define sigma-finite and even infinite measures and handle extended-real-valued measurable functions. It also allows us to prove the properties of the Lebesgue integral and its convergence theorems for arbitrary functions. We present a higher-order-logic formalization of the extended-real numbers in the following section.

3.1 Formalization of Extended Real Numbers

The set of extended-real numbers $\overline{\mathbb{R}}$ is the set of real numbers $\mathbb{R}$ extended with two additional elements, namely, the positive infinity $+\infty$ and negative infinity $-\infty$. $\overline{\mathbb{R}}$ is useful to describe various limiting behaviors in many mathematical fields. For instance, it is necessary to use the extended reals system to define the integration theory, otherwise the convergence theorems such as the monotone convergence and dominated convergence theorems would be less useful. Using the extended reals to define the measure theory makes it possible to define sigma-finite measures and other infinite measures. With extended-real numbers, the limit of a monotonic sequence is always defined, infinite when the sequence is divergent, but still defined and properties can be proven on it. The price to pay for these advantages is an increased level of difficulty in the analysis and the need to prove a large body of theorems on the extended reals and operators on them.

3.1.1 Type and Operators

An extended real is either a normal real number, positive infinity or negative infinity. we use `Hol_datatype` to define the new type `extreal` as follows:

```
val _ = Hol_datatype ‘extreal = NegInf | PosInf | Normal of real‘
```

The HOL notation `x:real` is used to specify the type of the variable x, which, in this case, is the `real` type for real numbers. After defining the new type, `Normal (1:real)` is now of type `:extreal`. To simplify the notation and represent `Normal 1` as `1`, which is in this case an extended real number (`1:extreal`), we use the following functions:

```
val extreal_of_num_def = Define ‘extreal_of_num n = Normal (&n)‘;
val _ = add_numeral_form (#"x", SOME "extreal_of_num");
```

All arithmetic operations of the real numbers need to be extended for the new type. To do so, we define HOL functions over the new type and then overload the common operators using these functions. For instance, we define the addition operation over the extended-real numbers using the function `extreal_add` presented below, and overload the `+` operator with this function. The function `extreal_add` extends the addition as follows:

$$\forall a.\ a \neq -\infty \ \Rightarrow\ a + (+\infty) \ = \ +\infty + a \ = \ +\infty$$

$$\forall a.\ a \neq +\infty \ \Rightarrow\ a + (-\infty) \ = \ -\infty + a \ = \ -\infty$$

This is formalized in higher-order logic as:

```
⊢ extreal_add (Normal x) (Normal y) = Normal (x + y) ∧
  extreal_add (Normal _) a = a ∧
  extreal_add b (Normal _) = b ∧
  extreal_add NegInf NegInf = NegInf ∧
  extreal_add PosInf PosInf = PosInf
```

The function is left undefined when one of the operands is `PosInf` and the other is `NegInf`. The + operator is then overloaded as

```
val _ = overload_on ("+",   Term ‘extreal_add’);
```

Similarly, we extend the other arithmetic operators and prove their properties.

```
val _ = overload_on ("-",   Term ‘extreal_sub‘);
val _ = overload_on ("*",   Term ‘extreal_mul‘);
val _ = overload_on ("/",   Term ‘extreal_div‘);
val _ = overload_on ("≤",  Term ‘extreal_le‘);
val _ = overload_on ("<",   Term ‘extreal_lt‘);
val _ = overload_on ("~",   Term ‘extreal_ainv‘);
val _ = overload_on ("numeric_negate",   Term ‘extreal_ainv‘);
val _ = overload_on ("inv", Term ‘extreal_inv‘);
val _ = overload_on ("abs",   Term ‘extreal_abs‘);
val _ = overload_on ("logr",   Term ‘extreal_logr‘);
val _ = overload_on ("lg",   Term ‘extreal_lg‘);
val _ = overload_on ("exp",   Term ‘extreal_exp‘);
val _ = overload_on ("pow", Term ‘extreal_pow‘);
val _ = overload_on ("sqrt", Term ‘extreal_sqrt‘);
```

The order relation, for example, is extended as: $\forall a \in \overline{\mathbb{R}},\ -\infty \leq a \leq +\infty$.

```
⊢  extreal_le (Normal x) (Normal y) = (x ≤ y) ∧
   extreal_le NegInf a = T ∧
   extreal_le b PosInf = T ∧
   extreal_le c NegInf = F ∧
   extreal_le PosInf d = F
```

With this order, $\overline{\mathbb{R}}$ is a complete lattice where every subset has a supremum and an infimum.

3.1.2 Supremum and Infimum

The supremum or least upper bound of a set $s \subseteq \overline{\mathbb{R}}$ is the least element of $\overline{\mathbb{R}}$ that it is greater than or equal to every element of s. We formalize the supremum in HOL as:

```
⊢  extreal_sup p =
   if ∀x. (∀y. p y ⇒ y ≤ x) ⇒ (x = PosInf) then PosInf
   else (if ∀x. p x ⇒ (x = NegInf) then NegInf
               else Normal (sup (λr. p (Normal r))))’;
```

In this definition, `sup` refers to the supremum over a set of real numbers.

The infimum or greatest lower bound of a set $s \subseteq \overline{\mathbb{R}}$ is the greatest element of $\overline{\mathbb{R}}$ that it is less than or equal to every element of s. We use the definition of supremum to formalize the infimum as:

```
⊢  extreal_inf p = - extreal_sup (IMAGE numeric_negate p)
```

The function `numeric_negate` takes a set as an argument and returns a new set where all the elements of the original set are negated. We overload the `sup` and `inf` operators with these new definitions.

```
val _ = overload_on ("sup", Term ‘extreal_sup‘);
val _ = overload_on ("inf", Term ‘extreal_inf‘);
```

Next, we prove the following theorem in HOL, which we will use in the Radon Nikodym theorem proof of Section 4.2.1.

Theorem 3.1. *For any non-empty, upper bounded (by a finite number) set P of extended real numbers, there exists a monotonically increasing sequence of elements of P that converges to the supremum of P.*

Proof. For the case where the supremum is an element of the set, we simply consider the sequence $\forall n$, $x_p(n) = \sup P$. Otherwise, we prove that $x_p(n)$, defined below, is one such sequence.

$$x_p(0) \;=\; @r.\ r \in P \;\wedge\; (\sup P - 1) < r \text{ and}$$

$$x_p(n+1) = @r.\ r \in P \;\wedge\; \max(x_p(n), \sup P - \tfrac{1}{2^{n+1}}) < r < \sup P$$

$$\vdash x_p\texttt{(0)} \;\; \texttt{=} \;\; \texttt{@r. r}\in\texttt{P} \wedge \texttt{(sup P - 1) < r}$$

$$\vdash x_p\texttt{(n+1) = @r. r}\in\texttt{P} \wedge \texttt{max(}x_p\texttt{(n), sup P - }\tfrac{1}{2^{n+1}}\texttt{) < r < sup P}$$

where `@` represents the Hilbert choice operator. □

3.1.3 Summation over a Set

We then define the sum of extended real numbers over a finite set and prove its properties whenever the sum is defined. The obvious way to define the sum is the following, where `ITSET` is the HOL function to iterate over sets,

```
val SIGMA_DEF = new_definition("SIGMA_DEF",
  ‘‘SIGMA f s = ITSET (λe acc. f e + acc) s (0:extreal)’’)
```

However, using this definition, we are not able to prove the recursive form without requiring that all the elements we are adding are finite. In fact, to be able to prove the recursive form, we need to use the theorem

```
 ∀f e s b.
   (∀x y z. f x (f y z) = f y (f x z)) ∧ FINITE s ⇒
   (ITSET f (e INSERT s) b = f e (ITSET f (s DELETE e) b))
```

This requires that the addition is associative and commutative for all the elements considered, which is not the case unless we restrict our definition to finite values.

This is, obviously, undesirable when working with extended-real numbers. Instead, we propose the following definition for the sum.

```
val SIGMA_def = let open TotalDefn
 in tDefine "SIGMA"
    ‘SIGMA (f:’a -> extreal) (s: ’a -> bool) =
       if FINITE s then
          if s={} then 0:extreal
          else f (CHOICE s) + SIGMA f (REST s)
       else ARB‘
  (WF_REL_TAC ‘measure (CARD o SND)‘ THEN
   METIS_TAC [CARD_PSUBSET, REST_PSUBSET])
 end;
```

We use WF_REL_TAC to initiate the termination proof of the definition with the measure function measure (CARD o SND). We also use the first-order decision procedure METIS_TAC which we apply to the two theorems CARD_PSUBSET and REST_PSUBSET stating that REST s is a proper subset of s and that the cardinal of a proper subset of a set s is smaller than that of s. The functions CHOICE and REST return some element of a set and the remaining subset obtained by excluding that element from the original set, respectively.

From this definition, we prove the recursive form, which will be used in proving the main properties of the sum.

```
∀f s. FINITE s  ⇒
   ∀e. (∀x. x ∈ e INSERT s ⇒ f x ≠ NegInf) ∨
       (∀x. x ∈ e INSERT s ⇒ f x ≠ PosInf) ⇒
    (SIGMA f (e INSERT s) = f e + SIGMA f (s DELETE e))
```

Notice that we can have infinite values as long as the sum is defined. The properties that we proved include the linearity, monotonicity, and the summation over disjoint sets and products of sets.

Finally, we define the infinite sum of extended real numbers $\sum_{n \in \mathbb{N}} x_n$ using the `SIGMA` and `sup` operators and prove its properties.

```
val ext_suminf_def = Define
   ‘ext_suminf f = sup (IMAGE (λn. SIGMA f (count n)) UNIV)’
```

We provided an extensive formalization of the extended real numbers, which consists of more than 220 theorems written in around 4500 lines of code. It contains all the necessary tools to formalize most of the concepts that we need in measure, integration, probability and information theories. In the next sections, we present the formalization of these theories based on the extended-real numbers.

3.2 Formalization of Measure Theory in HOL

A measure is a way to assign a number to a set, interpreted as its size, and can be considered as a generalization of the concepts of length, area, volume, etc. Two important examples are the Lebesgue measure on a Euclidean space and the probability measure on a Borel space. The former assigns the conventional length, area and volume of Euclidean geometry to suitable subsets of $\mathbb{R}^n, n = 1, 2, 3$ and the latter assigns a probability to an event and satisfies the condition that the measure of the sample space is equal to 1.

We provide a formalization of measure theory based on the Zermelo-Fraenkel set theory [21] with the famous Axiom of Choice (ZFC). This set theory is the most common foundation of mathematics up to the present day and allows to avoid a number of paradoxes caused by the use of the naive set theory. The Axiom of Choice, however, implies the existence of counter-intuitive sets and gives rise to paradoxes of its own, in particular, the Banach-Tarski paradox [67], which says that it is possible to decompose a solid unit ball into finitely many pieces and reassemble them into two copies of the original ball, using only rotations and no scaling. This paradox shows that there is no way to define the volume in three dimensions in the context of the ZFC set theory and at the same time requires that the rotation preserves the volume, and that the volume of two disjoint sets is the sum of their volumes. The solution to this is to tag some sets as non-measurable and to assign a volume only to a measurable set. Consequently, a measure function is defined over a class of subsets called the measurable sets and assigns a non-negative real number to every measurable set. It satisfies the countable additivity condition which states that the measure of the union of a collection of disjoint sets is equal to the sum of their measures.

3.2.1 Basic Definitions

Let $\mathcal{A}$ be a collection of subsets (or subset class) of a space X. We define a predicate `subset_class` in HOL that will test whether $\mathcal{A}$ is a subset class of X. This is formalized in HOL as:

```
⊢ subset_class X A = ∀s. s ∈ A ⇒ s ⊆ X
```

We also define the `countable` function that will test whether a set s is countable. A set is countable if its elements can be counted one at a time, or in other words, if there exists a surjective function $f : \mathbb{N} \to s$ such that every element of the set s can be associated with a natural number. This is formalized in HOL as:

```
⊢ countable s = ∃f. ∀x. x ∈ s ⇒ ∃(n:num). f n = x
```

Definition 3.1. *(Sigma Algebra)*

Let $\mathcal{A}$ be a collection of subsets (or subset class) of a space X. $\mathcal{A}$ defines a sigma algebra on X iff *$\mathcal{A}$ contains the empty set $\emptyset$, and is closed under countable unions and complementation within the space X.*

We formalize a sigma algebra in HOL as follows:

```
⊢ sigma_algebra (X,A) =
    subset_class X A ∧ {} ∈ A ∧
    (∀s. s ∈ A ⇒ X\s ∈ A) ∧
    ∀c. countable c ∧ c ⊆ A ⇒ ⋃c ∈ A
```

where $X \backslash s$ denotes the complement of s within X and $\bigcup c$ the union of all elements of c.

The pair $(X, \mathcal{A})$ is called a σ-field or a measurable space and $\mathcal{A}$ is the set of measurable sets. We define the `space` and `subsets` functions such that

```
⊢ space (X,A) = X
⊢ subsets (X,A) = A
```

Trivial examples of a sigma algebra on a space X include the empty set, which is the smallest sigma algebra on X and the powerset of X, $\mathcal{P}(X)$ which is comprised of all subsets of X and is the largest sigma algebra on X.

For any collection G of subsets of X, we can construct $\sigma(X, G)$, the smallest sigma algebra on X containing G. $\sigma(X, G)$ is called the sigma algebra on X generated by G. There is at least one sigma algebra on X containing G, namely the powerset of X. $\sigma(X, G)$ is the intersection of all those sigma algebras. The sigma algebra on X generated by G is formalized in HOL as:

```
⊢ sigma X G = (X, ⋂{s | G ⊆ s ∧ sigma_algebra (X,s)})
```

where $\bigcap c$ denotes the intersection of all elements of c.

Definition 3.2. *(Measure Space)*
A triplet $(X, \mathcal{A}, \mu)$ is a measure space iff *$(X, \mathcal{A})$ is a measurable space and $\mu : \mathcal{A} \to \overline{\mathbb{R}}$ is a non-negative and countably additive measure function.*

A probability space $(\Omega, \mathcal{A}, p)$ is a measure space satisfying $p(\Omega) = 1$. We formalize a measure space in HOL as follows:

```
⊢ measure_space (X,A,μ) =
    sigma_algebra (X,A) ∧ positive (X,A,μ) ∧
    countably_additive (X,A,μ)
```

A measure function is countably additive when the measure of a countable union of pairwise disjoint measurable sets is the sum of their respective measures. The countable additivity property is formalized in HOL as:

```
⊢ countably_additive (X,A,μ) =
    ∀f. f ∈ (UNIV → A) ∧
      (∀m n. m ≠ n ⇒ DISJOINT (f m) (f n)) ∧
      ⋃ (IMAGE f UNIV) ∈ A ⇒
      μ o f sums μ(⋃(IMAGE f UNIV))
```

In this definition, the countable union of subsets is captured through the domain of the function f which is the set of natural numbers `UNIV(:num)` and the range of f which is the set of measurable sets $\mathcal{A}$. The function μ is then countably additive if the sequence $\mu(f(n))$ converges to $\mu(\bigcup_n f(n))$.

We define the helper functions `m_space`, `measurable_sets` and `measure` which take a measure space as an argument and return the correspond component as follows:

```
⊢ m_space (X,A,μ) = X
⊢ measurable_sets (X,A,μ) = A
⊢ measure (X,A,μ) = μ
```

There is a special class of functions, called measurable functions, that are structure preserving, in the sense that the inverse image of each measurable set is also measurable. This is analogous to continuous functions in metric spaces where the inverse image of an open set is open. Measurable functions will be used in the next sections to define random variables.

Definition 3.3. *(Measurable Functions)*
Let $(X_1, \mathcal{A}_1)$ *and* $(X_2, \mathcal{A}_2)$ *be two measurable spaces. A function* $f : X_1 \rightarrow X_2$ *is called measurable with respect to* $(\mathcal{A}_1, \mathcal{A}_2)$ *(or* $(\mathcal{A}_1, \mathcal{A}_2)$ *measurable)* iff $f^{-1}(A) \in \mathcal{A}_1$ *for all* $A \in \mathcal{A}_2$*.*

The HOL formalization is the following:

```
⊢ f ∈ measurable a b =
    sigma_algebra a ∧ sigma_algebra b ∧ f ∈ (space a → space b) ∧
    ∀s. s ∈ subsets b ⇒ PREIMAGE f s ∩ space a ∈ subsets a
```

The HOL function `PREIMAGE` denotes the inverse image of a function. Notice that unlike Definition 3.3, the inverse image in the formalization needs to be intersected with `space a` because the functions in HOL are total, meaning that they map every value of a certain HOL type (even those outside `space a`) to a value of an appropriate type which may or may not be in `space b`. In other words, writing in HOL that f is a function from `space a` to `space b`, does not exclude values outside `space a` and hence the intersection is needed.

In this definition, we did not specify any structure on the measurable spaces. If we consider a function f that takes its values on a metric space, most commonly the set of real numbers or complex numbers, then the Borel sigma algebra on that space is used. In the following, we present our formalization of the Borel sigma algebra in HOL.

3.2.2 Borel Sigma Algebra

Working with the Borel sigma algebra makes the set of measurable functions a vector space. It also allows us to prove various properties of the measurable functions necessary for the formalization of the Lebesgue integral and prove its properties in HOL. The Borel sigma algebra on a space X is the smallest sigma algebra generated by the open sets of X. We use the `sigma` function we defined earlier to formalize the Borel sigma algebra.

```
⊢ borel X = sigma X (open_sets X)
```

An important example, especially in the theory of probability, is the Borel sigma algebra on $\overline{\mathbb{R}}$, denoted by $\mathcal{B}(\overline{\mathbb{R}})$ which we simply call *Borel* in the sequel.

```
⊢ Borel = sigma UNIV(:extreal) (open_sets UNIV)
```

where `UNIV` is the universal set of extended real numbers $\overline{\mathbb{R}}$. Clearly, the formalization of the Borel sigma algebra, which is based on the open sets, requires the formalization of some topology concepts of $\overline{\mathbb{R}}$. An earlier theory of the topology of $\mathbb{R}$ has been developed in HOL by Harrison [28]. Unfortunately, it does not use the set theory and also lacks some of the important theorems that we need in our development. Harrison, later, developed an extensive topology theory [29] in HOL-Light. Additionally, a formalization of the set of rational numbers $\mathbb{Q}$ is need in our work to prove, for instance, the various properties of $\mathcal{B}(\overline{\mathbb{R}})$. A theory for the rational numbers was also developed in HOL but does not include the theorems that we need and is in fact unusable for our development because we need to work with rational numbers as a subset of real numbers and not of a different HOL type.

Rational Numbers

A rational number is any number that can be expressed as the quotient of two integers, the denominator of which is positive. We use natural numbers and express $\mathbb{Q}$, the set of rational numbers, as the union of non-negative ($\mathbb{Q}^+$) and non-positive ($\mathbb{Q}^-$) rational numbers.

$$\vdash \mathbb{Q} = \{r \mid \exists\ n, m.\ r = \tfrac{n}{m} \ \wedge\ m > 0\} \ \cup\ \{r \mid \exists\ n, m.\ r = \tfrac{-n}{m} \ \wedge\ m > 0\}$$

We prove in HOL an extensive number of reassuring properties on the set $\mathbb{Q}$ as well as a few other less straightforward ones, namely, $\mathbb{Q}$ is countable, infinite and dense in $\overline{\mathbb{R}}$.

Theorem 3.2. $\mathbb{N} \subset \mathbb{Q}$ *and* $\forall x, y \in \mathbb{Q},\ -x,\ x+y,\ x-y,\ x*y\ \in \mathbb{Q}$ *and* $\forall y \neq 0,\ \frac{1}{y}$ *and* $\frac{x}{y}\ \in \mathbb{Q}$

A proof of this theorem in HOL is at the same time straightforward and tedious but it is necessary to manipulate elements of the newly defined set of rational numbers and prove their membership to $\mathbb{Q}$ in the following theorems.

Theorem 3.3. *The set of rational numbers* $\mathbb{Q}$ *is countable.*

Proof. We prove that there exists a bijection $f_1 : \mathbb{N} \to \mathbb{N} \times \mathbb{N}^*$ from the set of natural numbers $\mathbb{N}$ to the cross product of $\mathbb{N}$ and the set of positive natural numbers $\mathbb{N}^*$. Let $f_2 : \mathbb{N} \times \mathbb{N}^* \to \mathbb{Q}^+$ such that $f_2(a, b) = \frac{a}{b}$. and $f = f_2 \circ f_1$. Then $\forall x \in \mathbb{Q}^+$, there exists

$n \in \mathbb{N}$ such that $f(n) = x$. This proves that $\mathbb{Q}^+$ is countable. Similarly, we prove that $\mathbb{Q}^-$ is countable and that the union of two countable sets is countable. □

Theorem 3.4. *($\mathbb{Q}$ dense in $\overline{\mathbb{R}}$)*
$\forall x, y \in \overline{\mathbb{R}}$ and $x < y$, there exists $r \in \mathbb{Q}$ such that $x < r < y$.

Proof. We start by defining the ceiling of x as the smallest natural number larger than x, denoted by $\lceil x \rceil$ and prove that $\forall x,\ x \leq \lceil x \rceil$ and $\forall x \geq 0, \lceil x \rceil < x + 1$. Let $x, y \in \overline{\mathbb{R}}$ such that $x < y$. We use the ceiling function and the Archimedean property to construct r such that $x < r < y$. □

Another definition that will be useful in our development is the set of open intervals with rational end-points $I_r = \{]r_1, r_2[: r_1, r_2 \in \mathbb{Q}\}$.

```
⊢ open_intervals_set = {{x | a<x ∧ x<b} | a ∈ UNIV ∧ b ∈ UNIV}
```

We prove that I_r is countable by showing that the mapping $I_r \to \mathbb{Q} \times \mathbb{Q}$ that sends an open interval $]r_1, r_2[\in I_r$ to the ordered pair of rational numbers $(r_1, r_2) \in \mathbb{Q} \times \mathbb{Q}$ is injective, and that the cross product of two countable sets, $\mathbb{Q}$ in this case, is countable.

Topology

To define the Borel sigma algebra on $\overline{\mathbb{R}}$, we need some concepts of the topology of $\overline{\mathbb{R}}$ formalized in HOL. We could not use the formalization of topology by Harrison [28] because it does not support extended real numbers, it does not use the set theory and also lacks some of the important theorems that we need in our development. In the following, we define the concepts of neighborhood and open set in $\overline{\mathbb{R}}$ and prove the required theorems.

Definition 3.4. *Let $a \in A \subset \overline{\mathbb{R}}$. A is a neighborhood of a* iff *there exists a real number $d > 0$ such that $\forall x.\ |x - a| < d \Rightarrow\ x \in A$. In other words, a is an interior point of A.*

```
⊢ neighborhood A a = ∃d. 0<d ∧ ∀y. a - d < y ∧ y < a + d ⇒ y ∈ A
```

Definition 3.5. *A set that is a neighborhood to all of its points in an open set. Equivalently, if every point of a set is an interior point then the set is open.*

```
⊢ ∀A. open_set A = ∀x. x ∈ A ⇒ neighborhood A x
```

The following are some of the several properties related to open sets that we proved in HOL.

Theorem 3.5. *(Open Sets Properties)*

Property 1: The empty set and the universal set are open.

Property 2: Every open interval is an open set.

Property 3: The union of any family of open sets is open.

Property 4: The intersection of a finite number of open sets is open.

Property 5: Every open set in $\overline{\mathbb{R}}$ is the union of a countable family of open intervals.

Proof. We only show the proof for Property 5. Let A be an open set in $\overline{\mathbb{R}}$, then by the definition of open set, for all x in A there exists an open interval containing x such that $]a, b[\subset A$. Using the property of density of $\mathbb{Q}$ in $\overline{\mathbb{R}}$, there exists $]a_r, b_r[\subset A$ containing x, a_r and b_r being rational numbers. A is the union of a family of elements of I_r which is then countable because I_r is countable. □

Theorem 3.6. *The inverse image of an open set by a continuous function is open.*

Proof. Let A be an open set in $\overline{\mathbb{R}}$. From the previous theorem, A is a countable union of open intervals (A_i). $f^{-1}(A) = f^{-1}(\bigcup A_i) = \bigcup f^{-1}(A_i)$. Using Property 3, it suffices to prove that the inverse image of an open interval is open. For this we use the definition of a continuous function and the limit of a function to prove that any point of $f^{-1}(A_i)$ is an interior point. □

Borel Measurable Sets

We prove in this section that the Borel sigma algebra on the real line $\mathcal{B}(\overline{\mathbb{R}})$ is generated by the open intervals ($]c, d[$ for $c, d \in \overline{\mathbb{R}}$). This is actually used in many textbooks as a starting definition for the Borel sigma algebra on $\overline{\mathbb{R}}$. While we will prove that the two definitions are equivalent in the case of the real line, our formalization is vastly more general and can be used for any metric space such as the complex numbers or $\overline{\mathbb{R}}^n$, the n-dimensional Euclidian space.

Theorem 3.7. *$\mathcal{B}(\overline{\mathbb{R}})$ is generated by the open intervals $]c, d[$ where $c, d \in \overline{\mathbb{R}}$*

```
⊢ Borel = sigma UNIV (open_intervals_set)
```

Proof. The sigma algebra generated by the open intervals, σ_I, is by definition the intersection of all sigma algebras containing the open intervals. $\mathcal{B}(\overline{\mathbb{R}})$ is one of them because the open intervals are open sets (Property 2). Hence, $\sigma_I \subseteq \mathcal{B}(\overline{\mathbb{R}})$. Conversely, $\mathcal{B}(\overline{\mathbb{R}})$ is the intersection of all sigma algebras containing the open sets. σ_I is one of them because every open set on the real line is the union of a countable collection of open intervals (Property 5). Consequently $\mathcal{B}(\overline{\mathbb{R}}) \subseteq \sigma_I$ and finally $\mathcal{B}(\overline{\mathbb{R}}) = \sigma_I$. □

We also prove in HOL that $\mathcal{B}(\overline{\mathbb{R}})$ is generated by any of the following classes of intervals: $]-\infty, c[$, $[c, +\infty[$, $]c, +\infty[$, $]-\infty, c]$, $[c, d[$, $]c, d]$, $[c, d]$, where $c, d \in \overline{\mathbb{R}}$. To prove this result it suffices to prove that any interval $]c, d[$ is contained in the sigma algebra corresponding to each class. For the case of the intervals of type $[c, d[$, this follows from the equation $]c, d[= \bigcup_n [c + \frac{1}{2^n}, d[$. For the open rays $]-\infty, c[$, the result follows from the fact that $[a, b[$ can be written as the difference of two rays, $[a, b[=]-\infty, b[\setminus]-\infty, a[$. In a similar manner, we prove in HOL that all mentioned classes of intervals generate the Borel sigma algebra on $\overline{\mathbb{R}}$. Another useful result, asserts that the singleton sets are measurable sets of $\mathcal{B}(\overline{\mathbb{R}})$.

Theorem 3.8. $\forall c \in \overline{\mathbb{R}},\ \{c\} \in \mathcal{B}(\overline{\mathbb{R}})$

```
⊢ ∀c:real. {c} ∈ subsets Borel
```

The proof of this theorem follows from the fact that a sigma algebra is closed under countable intersections and the following equation.

$$\forall c \in \overline{\mathbb{R}} \quad \{c\} = \bigcap_n [c - \frac{1}{2^n}, c + \frac{1}{2^n}[$$

3.2.3 Extended-Real-Valued Measurable Functions

Recall that in order to check if a function f is measurable with respect to $(\mathcal{A}_1, \mathcal{A}_2)$, it is necessary to check that for any $A \in \mathcal{A}_2$, its inverse image $f^{-1}(A) \in \mathcal{A}_1$. The following theorem states that, for extended-real-valued functions, it suffices to perform the check on the open rays $((-\infty, c),\ c \in \overline{\mathbb{R}})$.

Theorem 3.9. *Let $(X, \mathcal{A})$ be a measurable space. A function $f : X \to \overline{\mathbb{R}}$ is measurable with respect to $(\mathcal{A}, \mathcal{B}(\overline{\mathbb{R}}))$ iff $\forall c \in \overline{\mathbb{R}},\ f^{-1}((-\infty, c)) \in \mathcal{A}$*

```
⊢ f ∈ measurable a Borel =
      sigma_algebra a ∧ f ∈ (space a → UNIV) ∧
      ∀c. {x | f x < c} ∩ space a ∈ subsets a
```

Proof. Suppose that f is measurable with respect to $(\mathcal{A}, \mathcal{B}(\overline{\mathbb{R}}))$, we showed in the previous section that $\forall c \in \overline{\mathbb{R}}$, $]-\infty, c[\in \mathcal{B}(\overline{\mathbb{R}})$. Since f is measurable then $f^{-1}(]-\infty, c[) \in \mathcal{A}$. Now suppose that $\forall c \in \overline{\mathbb{R}}$, $f^{-1}(]-\infty, c[) \in \mathcal{A}$, we need to prove $\forall\, A \in \mathcal{B}(\overline{\mathbb{R}})$, $f^{-1}(A) \in \mathcal{A}$. This follows from Property 5 stating that A is a countable union of open intervals and the equalities $f^{-1}(\bigcup_{n\in\mathbb{N}} A_n) = \bigcup_{n\in\mathbb{N}} f^{-1}(A_n)$ and $f^{-1}(]-\infty, c[) = \bigcup_{n\in\mathbb{N}} f^{-1}(]-n, c[)$ □

In a similar manner, we prove in HOL that f is measurable with respect to $(\mathcal{A}, \mathcal{B}(\overline{\mathbb{R}}))$ iff $\forall\, c, d \in \overline{\mathbb{R}}$ the inverse image of any of the following classes of intervals is an element of $\mathcal{A}$: $]-\infty, c[$, $[c, +\infty[$, $]c, +\infty[$, $]-\infty, c]$, $[c, d[$, $]c, d]$, $[c, d]$.

Every constant real function on a space X is measurable. The indicator function on a set A is measurable iff A is measurable.

In the following, we prove in HOL various properties of the real-valued measurable functions.

Theorem 3.10. *If f and g are $(\mathcal{A}, \mathcal{B}(\overline{\mathbb{R}}))$ measurable and $c \in \overline{\mathbb{R}}$ then cf, $|f|$, f^n, $f + g$, $f * g$ and $max(f, g)$ are $(\mathcal{A}, \mathcal{B}(\overline{\mathbb{R}}))$ measurable.*

```
⊢ ∀a f g h c.
    sigma_algebra a ∧
    f ∈ measurable a Borel ∧
    g ∈ measurable a Borel ⇒
       ((λx. c * f x) ∈ measurable a Borel)   ∧
       ((λx. abs(f x)) ∈ measurable a Borel)  ∧
       ((λx. f x pow n) ∈ measurable a Borel) ∧
       ((λx. f x + g x) ∈ measurable a Borel) ∧
       ((λx. f x * g x) ∈ measurable a Borel) ∧
       ((λx. max (f x) (g x)) ∈ measurable a Borel)
```

Theorem 3.11. *If (f_n) is a monotonically increasing sequence of extended-real-valued measurable functions with respect to $(\mathcal{A}, \mathcal{B}(\overline{\mathbb{R}}))$, such that $\forall\, x,\ f(x) = \sup_{n \in \mathbb{N}} f_n(x)$ then f is also $(\mathcal{A}, \mathcal{B}(\overline{\mathbb{R}}))$ measurable.*

```
 ⊢ ∀a f fi.
     sigma_algebra a ∧
     ∀i. fi i  ∈ measurable a Borel ∧
     ∀x. mono_increasing (λi. fi i x) ∧
     ∀x. x ∈ m_space m ⇒ f x = sup (IMAGE (λi. fi i x) UNIV)
             ⇒ f ∈ measurable a Borel
```

Theorem 3.12. *Every continuous function $g : \overline{\mathbb{R}} \to \overline{\mathbb{R}}$ is $(\mathcal{B}(\overline{\mathbb{R}}), \mathcal{B}(\overline{\mathbb{R}}))$ measurable.*

```
⊢ ∀g. (∀x. g contl x) ⇒ g ∈ measurable Borel Borel
```

Theorem 3.13. *If $g : \overline{\mathbb{R}} \to \overline{\mathbb{R}}$ is continuous and f is $(\mathcal{A}, \mathcal{B}(\overline{\mathbb{R}}))$ measurable then $g \circ f$ is also $(\mathcal{A}, \mathcal{B}(\overline{\mathbb{R}}))$ measurable.*

```
 ⊢ ∀a f g. sigma_algebra a ∧ f ∈ measurable a Borel ∧
       (∀x. g contl x) ⇒ g o f ∈ measurable a Borel
```

Theorem 3.12 is a direct result of Theorem 3.6 stating that the inverse image of an open set by a continuous function is open. Theorem 3.13 guarantees, for instance, that if f is measurable then $exp(f)$, $Log(f)$, $cos(f)$ are measurable. This is derived using Theorem 3.12 and the equality $(g \circ f)^{-1}(A) = f^{-1}(g^{-1}(A))$.

3.2.4 Products of Measure Spaces

We formalize products of measure spaces to be able to formalize measurable functions defined over product spaces. Let $m_1 = (X_1, \mathcal{S}_1, \mu_1)$ and $m_2 = (X_2, \mathcal{S}_2, \mu_2)$ be two measure spaces. The product of m_1 and m_2 is defined to be the measure space $(X_1 \times X_2, \mathcal{S}, \mu)$, where $\mathcal{S}$ is the sigma algebra on $X_1 \times X_2$ generated by subsets of the form $A_1 \times A_2$ where $A_1 \in \mathcal{S}_1$, and $A_2 \in \mathcal{S}_2$. The measure μ is defined for σ-finite measure spaces as

$$\mu(A) = \int_{X_1} \mu_2(\{y \in X_2 | (x,y) \in A\})\, d\mu_1$$

and $\mathcal{S}$ is defined using the `sigma` operator which returns the smallest sigma algebra containing a set of subsets, i.e., the product subsets in this case.

Let $g(s_1)$ be the function $s_2 \to (s_1, s_2)$ and `PREIMAGE` denote the HOL function for inverse image, then the product measure is formalized as

```
⊢ prod_measure m1 m2 =
    (λa. integral m1 (λs1. measure m2 (PREIMAGE g(s1) a)))
```

The integral in this definition is the Lebesgue integral for which we present the formalization in the next section. We verified in HOL that the product measure can be reduced to $\mu(a_1 \times a_2) = \mu_1(a_1) \times \mu_2(a_2)$ for finite measure spaces.

```
⊢ prod_measure m1 m2 (a1 × a2) = measure m1 a1 × measure m2 a2
```

We use the above definitions to define products of more than two measure spaces as follows. $X_1 \times X_2 \times X_3 = X_1 \times (X_2 \times X_3)$ and $\mu_1 \times \mu_2 \times \mu_3$ is defined as $\mu_1 \times (\mu_2 \times \mu_3)$. We also define the notion of absolutely continuous measures where μ_1 is said to be absolutely continuous w.r.t μ_2 iff for every measurable set A, $\mu_2(A) = 0$ implies $\mu_1(A) = 0$.

3.3 Formalization of Lebesgue Integration in HOL

Lebesgue integration [6] is a fundamental concept in many mathematical theories, such as real analysis [23], probability [26] and information, which are widely used to model and reason about the continuous and unpredictable components of physical systems. The reasons for its extensive usage, compared to the commonly known Riemann integral, include the ability to handle a broader class of functions, which are defined over more general types than the real line, and its better behavior when it comes to interchanging limits and integrals, which is of prime importance, for instance, in the study of Fourier series.

3.3.1 Lebesgue Integral

Similar to the way in which step functions are used in the development of the Riemann integral, the Lebesgue integral makes use of a special class of functions called positive simple functions. They are measurable functions taking finitely many values. In other words, a positive simple function g is represented by the triple (s, a, α) as a finite linear combination of indicator functions of measurable sets (a_i) that form a partition of the space X.

$$\forall x \in X,\ g(x) = \sum_{i \in s} \alpha_i I_{a_i}(x) \quad c_i \geq 0 \tag{3.1}$$

We also add the condition that positive simple functions take finite values, i.e., $\forall i \in s.\ x_i < \infty$. Their Lebesgue integral can however be infinite.
The Lebesgue integral is first defined for these functions and the definition is then extended to non-negative functions and finally to arbitrary functions.

Definition 3.6. *(Lebesgue Integral of Positive Simple Functions)*
Let $(X, \mathcal{A}, \mu)$ be a measure space. The integral of the positive simple function g with respect to the measure μ is defined as

$$\int_X g\, d\mu = \sum_{i \in s} \alpha_i \mu(a_i) \tag{3.2}$$

This is formalized in HOL as

```
⊢  pos_simple_fn_integral m s a α =
                SIGMA (λi. α_i * measure m (a i)) s
```

While the choice of $((\alpha_i), (a_i), s)$ to represent g is not unique, the integral as defined above is independent of that choice. Several properties of the Lebesgue integral of positive simple functions such as the linearity and monotonicity have been proven in [13]. We build on these results and extend them for our definition in which the Lebesgue integral is extended-real valued.

We use the definition of the Lebesgue integral of positive simple functions to define the integral of non-negative measurable functions using the supremum operator as follows:

$$\int_X f\,d\mu = \sup\{\int_X g\,d\mu \mid g \leq f \text{ and } g \text{ positive simple function}\} \tag{3.3}$$

Its formalization in HOL is the following:

```
⊢  pos_fn_integral m f =
           sup {r | ∃g. r ∈ psfis m g ∧ ∀x. g x ≤ f x}
```

where `psfis m g` is used to represent the Lebesgue integral of the positive simple function g. Finally, the integral for an arbitrary measurable function f is formalized in terms of the integrals of f^+ and f^- where f^+ and f^- are the non-negative functions defined by $f^+(x) = \max(f(x), 0)$ and $f^-(x) = \max(-f(x), 0)$.

$$\int_X f\,d\mu = \int_X f^+\,d\mu - \int_X f^-\,d\mu \tag{3.4}$$

Its formalization in HOL is the following.

```
⊢  integral m f = pos_fn_integral m (fn_plus f) -
                pos_fn_integral m (fn_minus f)
```

3.3.2 Lebesgue Monotone Convergence

The Lebesgue monotone convergence is arguably the most important theorem of the Lebesgue integration theory and it plays a major role in the proof of the Radon Nikodym theorem [9] and the properties of the integral. We present in the sequel a proof of the theorem in HOL.

Theorem 3.14. *Let (f_n) be a monotonically increasing sequence of non-negative measurable functions such that $\forall\, x,\ f(x) = \sup_{n \in \mathbb{N}} f_n(x)$, then*

$$\int_X f \, d\mu = \sup_{n \in \mathbb{N}} \int_X f_n \, d\mu$$

The higher-order-logic formalization of the Lebesgue monotone convergence is the following:

```
 ⊢ ∀m f fi. measure_space m ∧ ∀i x. 0 ≤ fi i x ∧
     ∀i. fi i ∈ measurable (m_space m, measurable_sets m) Borel ∧
     ∀x. mono_increasing (λi. fi i x) ∧
     ∀x. x ∈ m_space m ⇒ f x = sup (IMAGE (λi. fi i x) UNIV) ⇒
         pos_fn_integral m f =
             sup (IMAGE (λi. pos_fn_integral m (fi i)) UNIV)
```

We prove the Lebesgue monotone convergence theorem by using the properties of the supremum and by proving the lemma stating that if f is the supremum of a monotonically increasing sequence of non-negative measurable functions f_n and g is a positive simple function such that $g \leq f$, then the integral of g satisfies

$$\int_X g \, d\mu \leq \sup_{n \in \mathbb{N}} \int_X f_n \, d\mu$$

Or, as formalized in HOL:

```
⊢ ∀m f fi g r.
   measure_space m ∧ ∀i x. 0 ≤ fi i x ∧
   ∀i. fi i ∈ measurable (m_space m, measurable_sets m) Borel ∧
   ∀x. mono_increasing (λi. fi i x) ∧
   ∀x. x ∈ m_space m ⇒ (f x = sup (IMAGE (λi. fi i x) UNIV)) ∧
   r ∈ psfis m g ∧ ∀x. g x ≤ f x ⇒
       r  ≤ sup (IMAGE (λi. pos_fn_integral m (fi i)) UNIV)
```

3.3.3 Integrability

Our definition of the Lebesgue integral, based on the extended-real numbers, ensures that the integral is always defined for non-negative functions even when the integral is infinite. In this section, we define the criteria of integrability for an arbitrary measurable function and prove the integrability theorem which will play an important role in proving the properties of the Lebesgue integral.

Definition 3.7. *Let $(X, \mathcal{A}, \mu)$ be a measure space, a measurable function f is integrable iff $\int_X |f| \, d\mu < \infty$ or equivalently iff $\int_X f^+ \, d\mu < \infty$ and $\int_X f^- \, d\mu < \infty$*

```
⊢ integrable m f =
     f ∈ measurable (m_space m,measurable_sets m) Borel ∧
     pos_fn_integral m (fn_plus f) < ∞) ∧
     pos_fn_integral m (fn_minus f) < ∞)
```

We prove what we call the integrability theorem which has been used in some textbooks as a definition for integrability. This theorem provides also an alternative definition of the Lebesgue integral and plays an essential role to prove the properties of the Lebesgue Integral.

Theorem 3.15. *For any non-negative integrable function f there exists a sequence of positive simple functions (f_n) such that $\forall\, n, x,\ f_n(x) \leq f_{n+1}(x) \leq f(x)$ and $\forall\, x,\ f_n(x) \rightarrow f(x)$. In addition,*

$$\int_X f\, d\mu = \sup_{n \in \mathbb{N}} \int_X f_n\, d\mu$$

For arbitrary integrable functions, the theorem is applied to f^+ and f^- and results in a well-defined integral, given by

$$\int_X f\, d\mu = \sup_{n \in \mathbb{N}} \int_X f_n^+\, d\mu - \sup_{n \in \mathbb{N}} \int_X f_n^-\, d\mu$$

```
 ⊢ ∀m f. measure_space m ∧ integrable m f ⇒
     ∃fi ri.
       ∀x. mono_increasing (λi. fi i x) ∧
       ∀x. x ∈ m_space m ⇒
             (fn_plus f x = sup (IMAGE (λi. fi i x) UNIV)) ∧
       ∀i. ri i ∈ psfis m (fi i) ∧
       ∀i x. fi i x ≤ fn_plus f x ∧
       pos_fn_integral m (fn_plus f) = sup (IMAGE ri UNIV) ∧
     ∃gi vi.
       ∀x. mono_increasing (λi. gi i x) ∧
       ∀x. x ∈ m_space m ⇒
            (fn_minus f x = sup (IMAGE (λi. gi i x) UNIV)) ∧
       ∀i. vi i ∈ psfis m (gi i) ∧
       ∀i x. gi i x ≤ fn_minus f x ∧
       pos_fn_integral m (fn_minus f) = sup (IMAGE vi UNIV)
```

Proof. We prove the theorem by showing that the sequence (f_n), defined below, satisfies the conditions of the Lebesgue Monotone Convergence theorem. We apply this sequence for both f^+ and f^-.

$$f_n(x) = \sum_{k=0}^{4^n-1} \frac{k}{2^n} I_{\{x | \frac{k}{2^n} \leq f(x) < \frac{k+1}{2^n}\}} + 2^n I_{\{x | 2^n \leq f(x)\}} \tag{3.5}$$

First, we use the definition of (f_n) to prove in HOL the following lemmas

$\vdash$ ∀n x, 2^n $\leq$ f(x) $\Rightarrow$ f_n(x) = 2^n

$\vdash$ ∀n x and k < 4^n, $\frac{k}{2^n}$ $\leq$ f(x) < $\frac{k+1}{2^n}$ $\Rightarrow$ f_n(x) = $\frac{k}{2^n}$

$\vdash$ ∀x, (f(x) $\geq$ 2^n) $\vee$ (∃k, k < 4^n and $\frac{k}{2^n}$ $\leq$ f(x) < $\frac{k+1}{2^n}$)

Using these lemmas we prove that the sequence (f_n) is pointwise convergent to f $(\forall\, x,\ f(x) = \sup_n f_n(x))$, upper bounded by f $(\forall\, n, x,\ f_n(x) \leq f(x))$ and monotonically increasing $(\forall\, n, x,\ f_n(x) \leq f_{n+1}(x))$. □

3.3.4 Integral Properties

Most properties of the Lebesgue integral cannot be proved directly from the definition of the integral. Using the integrability theorem proven above, we can write the Lebesgue integral of any non-negative function as the supremum of sequence of integrals of positive simple functions. Using the properties of the supremum and the properties of the Lebesgue integral of positive simple functions, it is possible to prove the integral properties of non-negative functions and then for arbitrary measurable functions.

Let f and g be integrable functions and $c \in \mathbb{R}$ then

$\vdash\ \forall x,\ 0 \leq f(x) \Rightarrow 0 \leq \int_X f\, d\mu$

$\vdash\ \forall x,\ f(x) \leq g(x) \Rightarrow \int_X f\, d\mu \leq \int_X g\, d\mu$

$\vdash\ \int_X cf\, d\mu = c \int_X f\, d\mu$

$\vdash\ \int_X f + g\, d\mu = \int_X f\, d\mu + \int_X g\, d\mu$

$\vdash$ A and B disjoint sets $\Rightarrow \int_{A \cup B} f \, d\mu = \int_A f \, d\mu + \int_B f \, d\mu$

3.4 Summary and Discussions

We proposed in this chapter, a higher-order-logic formalization of the set of extended-real numbers which we used to formalize measure theory and Lebesgue integration in HOL. The formalization of measure theory includes the Borel sigma algebra allowing us to define measurable functions over arbitrary topological spaces and prove their properties. We formalized the Lebesgue Integral based on the extended-real numbers and we proved its main properties and convergence theorems. Both measure theory and Lebesgue integration formalization can be used to conduct the formal analysis of a wide range of engineering systems and protocols. We use them in the next section to formalize probability and information theory in HOL. The formalizations presented in this chapter required, approximatively, 15000 lines of codes, including definitions and theorems as well as their proofs. The formalization of extended-real numbers consists of more than 220 theorems. The numbers of theorems in the formalization of measure theory and Lebesgue integration are 120 and 100, respectively. It is worth mentioning that this number keeps decreasing starting from the formalization of the extended-real numbers and this can be explained by the way we are building the different theories on top of each other. This illustrates the advantage of the hierarchy we used to build the framework. The HOL scripts can be found in the official release of the HOL4 theorem prover [66].

Chapter 4

Formalization of Information Theory

In this chapter, we make use of the formalizations of measure theory and Lebesgue integration in HOL to provide a higher-order-logic formalization of probability theory. The latter is then used to formalize the main concepts of information theory. Finally, we give an overview of the quantitative analysis of information flow which is gaining lately a lot of interest in a wide range of applications, in particular, to evaluate the performance of anonymity and privacy protocols.

4.1 Formalization of Probability in HOL

Probability provides mathematical models for random phenomena and experiments. The purpose is to describe and predict relative frequencies (averages) of these experiments in terms of probabilities of events. The classical approach to formalize probabilities, which was the prevailing definition for many centuries, defines the probability of an event A as $p(A) = \frac{N_A}{N}$, where N_A is the number of outcomes favorable to the event A and N is the number of all possible outcomes of the experiment. Problems with this approach include the assumptions that all outcomes are equally likely (equiprobable) and that the number of possible outcomes is finite.

Kolmogorov [39] later introduced the axiomatic definition of probability, which provides a mathematically consistent way for assigning and deducing probabilities of events. This approach consists in defining a set of all possible outcomes, Ω, called the sample space, a set F of events which are subsets of Ω and a probability measure p such that (Ω, F, p) is a measure space with $p(\Omega) = 1$.

4.1.1 Basic Definitions

Definition 4.1. *(Ω, F, p) is a probability space if it is a measure space and $p(\Omega) = 1$.*

```
⊢ ∀p. prob_space p ⇔
    measure_space p ∧ (measure p (p_space p) = 1)
```

A probability measure is a measure function and an event is a measurable set.

```
⊢ prob = measure
⊢ events = measurable_sets
⊢ p_space = m_space
```

Definition 4.2. *Two events A and B are independent iff $p(A \cap B) = p(A)p(B)$.*

Here $A \cap B$ is the intersection of A and B, that is, it is the event that both events A and B occur.

```
⊢ independent p a b ⇔
     a ∈ events p ∧ b ∈ events p ∧
     prob p (a ∩ b) = prob p a * prob p b
```

Definition 4.3. *$X : \Omega \to \overline{\mathbb{R}}$ is a random variable iff X is $(F, \mathcal{B}(\overline{\mathbb{R}}))$ measurable*

where F denotes the set of events. Here we focus on real-valued random variables but the definition can be adapted for random variables having values on any topological space thanks to the general definition of the Borel sigma algebra.

```
⊢ random_variable X p s ⇔
      prob_space p ∧ X ∈ measurable (p_space p,events p) s
```

The properties we proved in the previous section for measurable functions are obviously valid for random variables.

Theorem 4.1. *If X and Y are random variables and $c \in \overline{\mathbb{R}}$ then the following functions are also random variables: $cX, |X|, X^n, X+Y, XY$ and $max(X,Y)$.*

Definition 4.4. *Two random variables X and Y are independent iff $\forall A, B \in \mathcal{B}(\overline{\mathbb{R}})$, the events $\{X \in A\}$ and $\{Y \in B\}$ are independent.*

The set $\{X \in A\}$ denotes the set of outcomes ω for which $X(\omega) \in A$. In other words $\{X \in A\} = X^{-1}(A)$.

```
⊢ independent_rv p X Y s t ⇔
      ∀A B. A ∈ subsets s ∧ B ∈ subsets t ⇒
      independent p (PREIMAGE X A ∩ p_space p) (PREIMAGE Y B ∩ p_space p)
```

The event $\{X \in A\}$ is used to define the probability mass function (PMF) of a random variable.

Definition 4.5. *The probability mass function p_X of a random variable X is defined as the function assigning to A the probability of the event $\{X \in A\}$.*

$$\forall A \in \mathcal{B}(\overline{\mathbb{R}}),\ p_X(A) = p(\{X \in A\}) = p(X^{-1}(A))$$

⊢ distribution p X = (λA. prob p (PREIMAGE X A ∩ p_space p))

The cumulative distribution function (CDF) of a random variable is defined as:

$$F_X(x) = P(X \leq x)$$

which we formalize in HOL as:

⊢ CDF p X = (λx. distribution p X {y | y ≤ x})

We formalize the joint distribution of two random variables as:

⊢ joint_distribution p X Y =

(λa. prob p (PREIMAGE (λx. (X x,Y x)) a ∩ p_space p))

The properties of the joint distribution that we proved in HOL include:

⊢ p_{XY}(a×b) = p_{YX}(b×a)

⊢ p_{XY}(a×b) ≤ p_X(a)

⊢ p_{XY}(a×b) ≤ p_Y(b)

⊢ FINITE (p_space p) ⇒ SIGMA (λ(x,y). p_{XY}{(x,y)}) (X(Ω)×Y(Ω)) = 1

⊢ FINITE (p_space p) ⇒ p_X(a) = SIGMA (λx. p_{XY}(a×{x})) Y(Ω)

⊢ FINITE (p_space p) ⇒ p_Y(b) = SIGMA (λx. p_{XY}({x}×b)) X(Ω)

⊢ FINITE (p_space p) ⇒

SIGMA (λ(x,y). p_{XY}{(x,y)} * f x) (X(Ω)×Y(Ω)) =

SIGMA (λx. p_X{x} * f x) X(Ω)

The conditional distribution of a random variable X given the random variable Y is formalized as:

⊢ conditional_distribution p X Y =

(λa b. joint_distribution p X Y (a b) / distribution p Y b)

4.1.2 Statistical Properties

In this section, we provide a formalization of the expectation, variance and covariance of a random variable in HOL. These definitions use the Lebesgue integration formalization and, hence, are valid for both discrete and continuous random variables.

Expectation

The expectation of a random variable is the weighted average of all its possible values.

Definition 4.6. *The expectation of a random value X is defined as the integral of X with respect to the probability measure.* $E[X] = \int_{\Omega} X \, dp$

```
⊢ expectation = integral
```

We prove the following properties of the expectation of random variables:

```
⊢  E[X+Y] = E[X] + E[Y]
⊢  E[aX] = aE[X]
⊢  E[a] = a
⊢  X ≤ Y then E[X] ≤ E[Y]
⊢  independent_rv X Y ⇒ E[XY] = E[X]E[Y]
```

As stated earlier, the definition of expectation is valid for both discrete and continuous cases. We prove for the case where the sample space is finite, the definition of the expectation can be simplied to:

```
 ⊢ FINITE (p_space p) ⇒
   E[X] = SIGMA (λr. r * distribution p X {r}) (IMAGE X (p_space p))
```

We provide a simplified definition for the expectation of real-valued random variables, which provides an easier way to compute the expecation of random variables without the need to work with integrals. This formula is derived from Equation 3.5 in Chapter 3.

$\vdash$ `E[X] =` $\sup_{n\in\mathbb{N}}\sum_{k=0}^{4^n-1}\frac{k}{2^n}\times(F_X(\frac{k+1}{2^n})-F_X(\frac{k}{2^n}))$ `+` $2^n\times(1-F_X(2^n))$

The conditional expectation is formalized in the following HOL definition:

$\vdash$ `conditional_expectation p X s =`

  `@f. real_random_variable f p` $\wedge$ `∀g. g` $\in$ `s` $\Rightarrow$

  `integral p (λx. f(x)`I_g`(x)) = integral p (λx. X(x)`I_g`(x))`

Variance

The variance is another descriptor of probability distributions providing a measure of how far the numbers are spread out around the expectation. The covariance is a measure of the correlation between two random variables.

Definition 4.7. *The variance of a random variable X is defined as $Var(X) = E[|X-E[X]|^2]$. The covariance of two random variables X and Y is defined as $Cov(X,Y) = E[(X-E[X])(Y-E[Y])]$. Two random variables X and Y are uncorrelated iff $Cov(X,Y) = 0$.*

Some of the properties that we verified in HOL for the variance and covariance include:

$\vdash$ `Var(X) =` $E[X^2]$ `-` $E[X]^2$

$\vdash$ `Cov(X,Y) = E[XY]-E[X]E[Y]`

$\vdash$ `Var(X)` $\geq$ `0`

$\vdash$ $\forall a\in\mathbb{R}$, `Var(`a`X) =` a^2`Var(X)`

$\vdash$ `Var(X+Y) = Var(X) + Var(Y) + 2Cov(X,Y)`

$\vdash$ `uncorrelated X Y` $\Rightarrow$ `Var(X+Y) = Var(X) + Var(Y)`

$\vdash$ $\forall$`i`$\neq$`j, uncorrelated` X_i X_j $\Rightarrow$ `Var(`$\sum_{i=1}^{N} X_i$`)=`$\sum_{i=1}^{N} Var(X_i)$

4.2 Formalization of Information Measures in HOL

Using the formalization of the foundational theories of measure, Lebesgue integration and probability, we are able to provide a higher-order-logic formalization of the main concepts of information theory. We start by formalizing the Radon-Nikodym derivative [23] and use it to define the KL divergence. The latter provides a unified framework based on which we define the most commonly used measures of information, those found in the main textbooks of Information Theory [14, 25, 22] such as the Shannon entropy and mutual information. We start by providing the general definitions which are valid for both discrete and continuous cases and then prove the corresponding reduced expressions where the measures considered are absolutely continuous over finite spaces. We build on the foundations, presented in [13], to provide a more general formalization of information theory including the properties of measures of information.

4.2.1 Radon-Nikodym Derivative

The Radon-Nikodym derivative of a measure ν with respect to the measure μ is defined as a non-negative measurable function f, satisfying the following formula, for any measurable set A [23].

$$\int_A f \, d\mu = \nu(A)$$

We formalize the Radon-Nikodym derivative in HOL as

```
⊢ RN_deriv m v =
    @f. f IN measurable (X,S) Borel ∧
    ∀x ∈ X, 0 ≤ f x ∧
    ∀a ∈ S, integral m (λx. f x × I_a x) = v a
```

where @ denotes the Hilbert-choice operator. The existence of the Radon-Nikodym derivative is guaranteed for absolutely continuous measures by the Radon-Nikodym theorem stating that if ν is absolutely continuous with respect to μ, then there exists a non-negative measurable function f such that for any measurable set A,

$$\int_A f \, d\mu = \nu(A)$$

We proved the Radon-Nikodym theorem in HOL for finite measures which can be easily generalized to σ-finite measures.

```
⊢ ∀m v s st.
    measure_space (s,st,m) ∧
    measure_space (s,st,v) ∧
    abs_cont (s,st,m) (s,st,v) ⇒
      ∃f. f ∈ measurable (s,st) Borel ∧
      ∀x ∈ s, 0 ≤ f x < ∞ ∧
      ∀a ∈ st, integral m (λx. f x × I_a x) = v a
```

The formal reasoning about the above theorem is primarily based on the Lebesgue monotone convergence and the following lemma which, to the best of our knowledge, has not been referred to in mathematical texts before

Lemma 1. *If P is a non-empty set of extended-real valued functions closed under the max operator, g is monotone over P and $g(P)$ is upper bounded, then there exists a* *monotonically increasing sequence $f(n)$ of functions, elements of P, such that:*

$$\sup_{n \in \mathbb{N}} g(f(n)) = \sup_{f \in P} g(f)$$

Proof. Proving the Radon Nikodym theorem consists in defining the set F of non-negative measurable functions such that for any measurable set A, $\int_A f \, d\mu \leq \nu(A)$. Then we prove that this set is non-empty, upper bounded by the finite measure of the space and is closed under the max operator. Next, using the monotonicity of the integral and the lemma above, we prove the existence of a monotonically increasing sequence $f(n)$ of functions in F such that:

$$\sup_{n \in \mathbb{N}} \int_X f_n \, d\mu = \sup_{f \in F} \int_X f \, d\mu$$

Finally, we define the function g such that $\forall x,\ g(x) = \sup_{n \in \mathbb{N}} f_n(x)$ and prove that g satisfies the conditions of the theorem by using the Lebesgue monotone convergence theorem to prove that:

$$\int_X g \, d\mu = \sup_{n \in \mathbb{N}} \int_X f_n \, d\mu$$

□

We formally verified various properties of the Radon-Nikodym derivative. For instance, we prove that for absolutely continuous measures defined over a finite space, the derivative reduces to

```
⊢ ∀x ∈ s, u{x} ≠ 0 ⇒
     RN_deriv u v x = v{x}/u{x}
```

The following properties play a vital role in formally reasoning about the Radon-Nikodym derivative and have also been formally verified.

```
⊢ ∀x ∈ s, 0 ≤ RN_deriv m v x < ∞
```

```
⊢ RN_deriv m v ∈ measurable (s,st) Borel
```

```
⊢ ∀a ∈ st, integral m (λx. RN_deriv m v x × I_a x) = v a
```

4.2.2 Kullback-Leibler Divergence

The Kullback-Leibler (KL) divergence [14] $D_{KL}(\mu||\nu)$ is a measure of the distance between two distributions μ and ν. It can be used to define most information-theoretic measures such as mutual information and entropy and can, hence, be used to provide a unified framework to formalize most information leakage measures. It is because of this reason that we propose to formalize the KL divergence in this paper as it will facilitate the formal reasoning about a wide variety of information flow related properties. The KL divergence is defined as:

$$D_{KL}(\mu||\nu) = -\int_X log \frac{d\nu}{d\mu}\, d\mu$$

where $\frac{d\nu}{d\mu}$ is the Radon-Nikodym derivative of ν with respect to μ. The KL divergence is formalized in HOL as:

```
⊢ KL_divergence b m v = -integral m (λx. log b (RN_deriv m v x))
```

where b is the base of the logarithm. D_{KL} is measured in *bits* when $b = 2$. We formally verify various properties of the KL divergence. For instance, we prove that for absolutely continuous measures over a finite space, it reduces to:

$$D_{KL}(\mu||\nu) = \sum_{x \in s} \mu\{x\} \log \frac{\mu\{x\}}{\nu\{x\}}$$

```
⊢ KL_divergence b u v = SIGMA (λx. u{x} log b (u{x} / v{x})) s
```

We also prove the following properties

```
⊢ KL_divergence b u u = 0
⊢ 1 ≤ b  ⇒  0 ≤ KL_divergence b u v
```

The non-negativity of the KL divergence for absolutely continuous probability measures over finite spaces is extensively used to prove the properties of information theory measures like mutual information and entropy. To prove this result, we use the Jensen's inequality [56] and the concavity of the logarithm function.

We show in the subsequent sections how we use the KL divergence to formalize mutual information, Shannon entropy, conditional entropy and the conditional mutual information, which are some of the most commonly used measures of information.

4.2.3 Mutual Information

The *mutal information* $I(X;Y)$ of two random variables is a measure of the mutual dependence of the two random variables in the sense that it measures how much uncertainty about one of these variables is reduced when the other variable is known. Mutual information has been proposed as a measure of information leakage [68] from the secure inputs X of a program to its public outputs Y. It is defined as the KL divergence between the joint distribution and the product of marginal distributions. The following is a formalization of mutual information in HOL.

```
⊢ I(X;Y) = KL_divergence b (joint_distribution p X Y)
                            prod_measure (distribution p X)
                                         (distribution p Y)
```

We prove various properties of mutual information in HOL, such as the non-negativity, symmetry and reduced expression for finite spaces, using the result that the joint distribution is absolutely continuous w.r.t the product of marginal distributions.

```
⊢ 0 ≤ I(X;Y)

⊢ I(X;Y) = I(Y;X)

⊢ I(X;Y) = 0 ⇔ independent_rv X Y

⊢ I(X;Y) = SIGMA (λ(x,y). p{(x,y)} log b (p{(x,y)}/p{x}p{y})) s
```

4.2.4 Shannon Entropy

The *Shannon entropy* $H(X)$ is a measure of the uncertainty associated with a random variable. Equivalently, it is a measure of the average information content missing when the value of the random variable is unknown. The Shannon entropy was one of the first measures to be proposed to analyze anonymity protocols and secure communications [61, 19]. It can be defined as the expectation of p_X or simply as $I(X;X)$.

```
⊢ H(X) = I(X;X)
```

We prove that it can also be expressed in terms of the KL divergence between p_X and the uniform distribution p_X^u, where N is the size of the alphabet of X.

```
⊢ H(X) = log(N) - KL_divergence b (distribution p X)
                                  (uniform_dist p X)
```

The *cross entropy* $H(X,Y)$ which measures how much entropy is contained in a joint system of two random variables is the entropy of the random variable (X,Y) and hence there is no need for a separate formalization of the cross entropy.

The *conditional entropy* $H(X|Y)$ quantifies the remaining uncertainty about the random variable X given that the value of the random variable Y is known. It is defined in terms of the KL divergence as follows:

```
⊢ H(X|Y) = log(N) - KL_divergence b (joint_distribution p X Y)
                               prod_measure (uniform_dist p X)
                                            (distribution p Y)
```

The entropy properties that we prove in HOL include:

```
⊢ 0 ≤ H(X) ≤ log(N)

⊢ max(H(X),H(Y)) ≤ H(X,Y) ≤ H(X) + H(Y)
⊢ H(X|Y) = H(X,Y) - H(Y)

⊢ 0 ≤ H(X|Y) ≤ H(X)

⊢ I(X;Y) = H(X) + H(Y) - H(X,Y)

⊢ I(X;Y)  ≤ min(H(X),H(Y))

⊢ H(X)   = -SIGMA (λx. p{x} log b (p{x})) s

⊢ H(X|Y) = -SIGMA (λ(x,y). p{(x,y)} log b (p{(x,y)}/p{y})) s
```

4.2.5 Conditional Mutual Information

The *conditional mutual information* $I(X;Y|Z)$ allows one to measure the expected value of mutual information of two random variables X and Y given knowledge of Z. It was used by Malacaria [44] as a measure of information leakage for a program with high security inputs X, low security outputs Y and low security inputs Z. $I(X;Y|Z)$ is then a measure of how much information about the secret inputs is leaked to an attacker by observing the outputs of a program given knowledge of the low security inputs. The conditional mutual information is defined as the KL divergence between the joint distribution p_{XYZ} and the product measure $p_{X|Z}p_{Y|Z}p_Z$. Its HOL formalization is as follows:

```
⊢ conditional_mutual_information b p X Y Z =
   KL_divergence b (joint_distribution p X Y Z)
                   (prod_measure (conditional_distribution p X Z)
                                 (conditional_distribution p Y Z)
                                 (distribution p Y))
```

We formally verify the following reduced form of the conditional mutual information for finite spaces by first proving that p_{XYZ} is absolutely continuous w.r.t $p_{X|Z}p_{Y|Z}p_Z$

and then apply the reduced form of the KL divergence.

$$I(X;Y|Z) = \sum_{(x,y,z)\in\mathcal{X}\times\mathcal{Y}\times\mathcal{Z}} p(x,y,z)\log\frac{p(x,y,z)}{p(x|z)p(y|z)p(z)}$$

When the two random variables X and Y are independent given Z, the conditional mutual information $I(X;Y|Z) = 0$. In fact, in this case,
$\forall x, y, z.\ p(x,y,z) = p(x,y|z)p(z) = p(x|z)p(y|z)p(z)$.

```
⊢ indep_rv_cond p X Y Z ⇒ I(X;Y|Z) = 0
```

We also prove a few other important results regarding the conditional mutual information which will be useful later in our work.

```
⊢ 0 ≤ I(X;Y|Z)

⊢ I(X;Y|Z) = H(X|Z) - H(X|Y,Z)

⊢ I(X;Y|Z) = I(X;(Y,Z)) - I(X;Z)

⊢ I(X;Y|Z) ≤ H(X|Z)
```

The first property is a direct result of the non-negativity of the KL divergence. We will show next a proof of the second property. In the same manner we prove the third property. Finally the fourth property is a result from the second property and the non-negativity of the entropy.

$$\begin{aligned}
\texttt{I(X;Y|Z)} \quad &= \quad \sum p(x,y,z)\log\tfrac{p(x,y,z)}{p(x|z)p(y|z)p(z)} \\
&= \quad \sum p(x,y,z)\log\tfrac{p(x|y,z)}{p(x|z)} \\
&= \quad \sum p(x,y,z)\log(p(x|y,z)) \texttt{ - } \sum p(x,z)\log(p(x|z)) \\
&= \quad \texttt{- H(X|Y,Z) + H(X|Z)}
\end{aligned}$$

So far, we have provided a higher-order-logic formalization of the KL divergence which we used to define various measures of quantitative information flow. This framework, along with the formalization of measure and probability theories, allows us to conduct many analyses of quantitative information flow using a theorem prover and hence guaranteeing the soundness of the analysis.

4.3 Quantitative Analysis of Information Flow

A classical approach to protecting the confidentiality of sensitive information is to use a noninterference technique which aims to make it independent of the public output of the protocol. Unfortunately, this cannot be applied to a large number of applications where a small leak of information is intended by design to ensure the functionality of the protocol. In an election protocol, for instance, while the votes should remain secret, the election results should be made public. Similarly, a password checker reveals some information when rejecting an incorrect password.

Quantitative analysis of information flow [64, 58] is gaining a lot of attention in a variety of contexts, such as secure information flow, anonymity protocols, and side-channel analysis. It allows to measure how much information about the high security inputs of a system can be leaked, accidentally or maliciously, by observing the systems outputs and possibly the low security inputs. Unlike non-interference analysis, which only determines whether a system is completely secure or not completely secure, quantitative information flow analysis provides an information-theoretic measure on how secure or insecure a system is.

Various measures are being proposed to quantify the flow of information. Serjantov [61] and Diaz et al. [19] independently proposed to use entropy to define the quality of anonymity and to compare different anonymity systems. In this technique, the attacker assigns probabilities to the users after observing the system and does not make use of any apriori information he/she might have. The attacker simply assumes a uniform distribution among the users before observation. Malacaria [44] defined the leakage of confidential information in a program as the conditional mutual information between its outputs and secret inputs, given the knowledge of its low security inputs. Chatzikokolakis [10] modeled anonymity protocols as noisy channels and used the channel capacity as a measure of the loss of anonymity. If it is equal to 0 then the attacker can learn nothing more by observing the protocol. In the cases where some leakage of information is intended by design as is the case in an election protocol, for example, another measure for the loss of anonymity, the conditional capacity, was proposed to take into account the intended leakage. In both cases however, there is no analytical formula to compute the capacity and numerical algorithms have to be used. Symmetry properties of channels, when present, can be exploited to compute the capacity. Deng [18] used the notion of relative entropy to measure the degree of anonymity that protocols can guarantee.

We introduce two new measures of information, namely the information leakage degree and the conditional information leakage degree, which can be used to evaluate the anonymity and security properties of various systems and protocols.

4.3.1 Information Leakage Degree

Consider a program having a set of secret inputs, represented by the random variable X and a set of public outputs, represented by Y. We define the information leakage degree of this program as

$$D = \frac{H(X|Y)}{H(X)}$$

where $H(X)$ and $H(X|Y)$ represent the Shannon entropy of X and the conditional entropy of X given Y, respectively.

```
⊢ D p X Y = conditional_entropy p X Y / entropy p X
```

To better understand the intuition behind this definition, let us consider the two extreme cases of a completely secure program and a completely insecure program. Complete security, intuitively, happens when the knowledge of the public output Y of a program does not affect the uncertainty about the secret input X. This is equivalent to the requirement that X is independent of Y. In this case $H(X|Y) = H(X)$ and the information leakage degree is equal to 1. On the other hand, when the output of the program completely identifies its secret input, the entropy $H(X|Y)$ is equal to 0 and hence the information leakage degree is equal to 0 in this case of perfect identification. For situations between the two extremes, we prove that the information leakage degree lies within the interval $(0, 1)$.

```
⊢ 0 ≤ D p X Y ≤ 1
```

Using the properties of mutual information, $I(X;Y)$, we prove that the information leakage degree is also equal to

$$D = 1 - \frac{I(X;Y)}{H(X)}$$

This result illustrates the significance of the information leakage degree definition since mutual information measures how much information an adversary can learn about the input X after observing the output Y. This also allows to compare our definition to the anonymity degree proposed in [68] as

$$D' = 1 - \frac{I(X;Y)}{logN}$$

where N is the size of the alphabet of X. Our definition is more general. In fact, when X is uniformly distributed, the two measures coincide $D = D'$. However, in the general case, we believe that our definition is more accurate since, for instance, in the perfect identification scenario, D is always equal to 1 regardless of the input distribution. On the other hand, D' is equal to 1 only in the special case of a uniform distribution. In [68] the authors considered using $H(X)$ as a normalization factor instead of $logN$ but opted for the latter arguing that the input distribution is already accounted for in mutual information. But as stated previously, with the definition of D', the proof for perfect identification is only valid for uniformly distributed inputs.

4.3.2 Conditional Information Leakage Degree

We propose another variation of information leakage degree that is more general and can cover a wider range of scenarios. First, consider a program which has a set of high security inputs S, a set of low security inputs L and a set of public outputs O. The adversary wants to learn about the high inputs S by observing the outputs O given the knowledge of the low inputs L. To capture this added information for the adversary (low inputs), we propose the following definition, which we call the conditional information leakage degree.

$$D_c = \frac{H(S|(O,L))}{H(S|L)}$$

This is formalized in HOL as

```
⊢ D_c p S L O =
          conditional_entropy p S (O,L) / conditional_entropy p S L
```

Just like the previous case, consider the two extremes of perfect security and perfect identification. When the outputs and the secret inputs are independent, given L, the conditional entropy $H(S|(O,L))$ is equal to $H(S|L)$ which results in a conditional leakage degree equal to 1 for perfect security. However, if the public inputs and outputs completely identify the secret inputs, then $H(S|(O,L))$ is equal to 0 and so is the conditional leakage degree in the case of perfect identification. As in the case of leakage degree, we are also able to show that the conditional information leakage degree lies within the interval $(0, 1)$.

⊢ 0 ≤ D_c p X Y Z ≤ 1

We also prove that the conditional information leakage degree can be written in terms of the conditional mutual information and the conditional entropy.

$$D_c = 1 - \frac{I(S;O|L)}{H(S|L)}$$

This shows that this definition is clearly a generalization of the information leakage degree for the case of programs with additional low security inputs.

We provide more intuition to interpret this definition by proving the *data processing inequality* (DPI) [14], which states that if the random variables X, Y and Z form a Markov chain, then $I(X;Z) \leq I(X;Y)$.

Definition 4.8. *Random variables X, Y, Z are said to form a Markov chain in that order (denoted by $X \rightarrow Y \rightarrow Z$) if the conditional distribution of Z depends only on Y and is conditionally independent of X. Specifically, X, Y and Z form a Markov chain $X \rightarrow Y \rightarrow Z$ if the joint probability mass function can be written as $p(x, y, z) = p(x)p(y|x)p(z|y)$.*

We formalize this in HOL as follows.

⊢ markov_chain p X Y Z ⇔
    ∀x y z, p_{XYZ}{(x,y,z)} = p_X{x} * $p_{Y|X}${y}{x} * $p_{Z|Y}${z}{y}

We prove that $X \rightarrow Y \rightarrow Z$ is equivalent to the statement that X and Z are conditionally independent given Y.

⊢ markov_chain p X Y Z ⇔ indep_rv_cond p X Z Y

In fact, $p(x)p(y|x)p(z|y) = p(x,y)p(z|y) = p(x|y)p(z|y)p(y)$. This in turn is equivalent to $I(X;Z|Y) = 0$.

```
⊢ markov_chain p X Y Z ⇔ I(X;Z|Y) = 0
```

This result allows us to prove the DPI as follows:

```
⊢ markov_chain p X Y Z ⇒ I(X;Z) ≤ I(X;Y)
```

We prove the DPI theorem using the properties of mutual information. In fact, as shown previously,

```
⊢ I(X;(Y,Z)) = I(X;Z) + I(X;Y|Z)
```

By symmetry of mutual information, we also have

```
⊢ I(X;(Y,Z)) = I(X;Y) + I(X;Z|Y)
```

Since $I(X;Z|Y) = 0$ for a Markov chain,

```
⊢ I(X;(Y,Z)) = I(X;Y)
```

Using the non-negativity of the conditional mutual information, it is straightforward to conclude that

```
⊢ I(X;Z) ≤ I(X;Y)
```

The data processing inequality is an important result in information theory that is used, for instance, in statistics to define the notion of sufficient statistic. We make use of the DPI to interpret the conditional information leakage degree. For a system with high security inputs S, low security inputs L and outputs O, if the outputs depend only on the low inputs, i.e., $p(O|S,L) = p(O|L)$ then $S \to L \to O$ and S and O are conditionally independent given L. This is the perfect security scenario, for which

$D_c = 1$. Using the DPI, we conclude that $I(S;O) \leq I(S;L)$. This means that when the conditional mutual information leakage is equal to 1, no clever manipulation of the low inputs, by the attacker, deterministic or random, can increase the information that L contains about S, $(I(S;L))$.

4.4 Summary and Discussions

In this chapter, we used the formalization of measure theory and Lebesgue information to provide a higher-order-logic formalization of the main concepts of probability and information theory. We formalized the basic definitions of probability spaces and random variables as well as their statistical properties. Based on the formalization of the KL divergence, we defined the most commonly used measures of information including the Shannon entropy and mutual information. We have also introduced two novel measures of information leakage that we will use in Chapter 5 to reason about information flow of real-world protocols and programs. The formalization of probability presented in this chapter required around 3000 lines of code, including basic definitions and proofs of 60 theorems. The formalization of information theory required 3000 lines of code and 40 theorems. This is a great indication of the decreasing trend of the number of theorems thanks to the hierarchy used within the framework. These formalizations are now part of the official release [47] of the HOL4 theorem prover.

Chapter 5

Applications

In previous chapters, we provided a comprehensive framework that can be used in the formal probabilistic and information-theoretic analysis of a wide range of systems and protocols. We illustrate the usefulness of conducting this analysis using theorem proving by tackling a number of applications including a data compression application, the formal analysis of an anonymity-based MIX channel and the properties of the one-time pad encryption system.

5.1 Data Compression

Data compression or source coding may be viewed as a branch of information theory in which the primary objective is to reduce redundancy, minimizing the amount of data to be stored or transmitted. It consists in encoding information using fewer bits than an unencoded representation would use, through use of specific encoding schemes. As depicted in Figure 5.1, data compression has important applications in the areas of data storage and data transmission, for instance, in the speech compression for real-time transmission over digital cellular networks. On the other hand, the proliferation

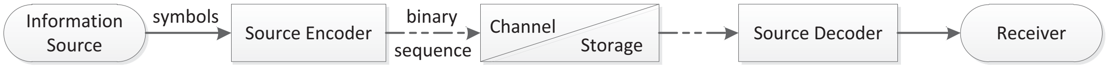

Figure 5.1: Source Coding

of computer communication networks is resulting in massive transfer of data and the growing number of data processing applications require storage of large volumes of data. Compressing data reduces costs and increases capacity. However, to guarantee reliable transmission or storage of data, particularly in systems for safety-critical applications, a certain limit of compression must be respected.

We propose to formally prove an important result establishing the fundamental limit of data compression. This result is also known as the Shannon source coding theorem and states that it is possible to compress the data at a rate that is arbitrarily close to the Shannon entropy without significant loss of information. In other words, n independent and identically distributed (iid) random variables with entropy $H(X)$ can be expressed on the average by $nH(X)$ bits without significant risk of information loss, as n tends to infinity. To the best of our knowledge, most of the formalization framework that we need for this application, such as the properties of real valued measurable functions, properties of the expectation of arbitrary functions, variance, independence of random variables and the weak law of large numbers, is not available in the open literature.

A proof of this result consists in proposing an encoding scheme for which the average codeword length can be made arbitrarily close to $nH(X)$ with negligible probability of loss. We propose to perform a *typical set* encoding [14]. The typical set contains the typical sequences which are used to represent the frequent source symbols while the non-typical sequences represent rare source symbols.

The encoding is performed separately for the typical and the non-typical set as

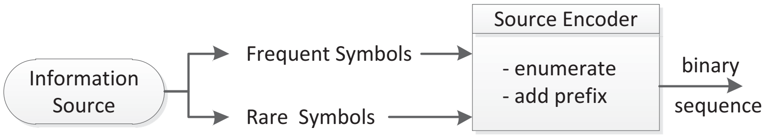

Figure 5.2: Typical Set Encoding

depicted in Figure 5.2. Clearly, we need to formalize the concept of a typical set and prove its properties to be able to formally verify that the average codeword length associated to the typical-set encoding can be made arbitrarily close to the Shannon entropy with a vanishing probability of error. To prove the properties of the typical set, we need to prove the *Asymptotic Equipartition Property* (AEP) [14, 25], which in turn, requires the proof of the classical probability results, namely, the Weak Law of Large Numbers (WLLN) and the Chebyshev inequality.

5.1.1 Chebyshev and Markov Inequalities

In probability theory, both the Chebyshev and Markov inequalities provide estimates of tail probabilities. The Chebyshev inequality guarantees, for any probability distribution, that nearly all the values are close to the mean and it plays a major role in the derivation of the laws of large numbers [49]. The Markov inequality provides loose yet useful bounds for the cumulative distribution function of a random variable. Let X be a random variable with expected value m and finite variance σ^2. The Chebyshev inequality states that for any real number $k > 0$,

$$P(|X - m| \geq k\sigma) \leq \frac{1}{k^2} \tag{5.1}$$

The HOL formalization of the Chebyshev inequality is stated as:

```
⊢ random_variable X p Borel ∧
    integrable p (λx. (Xx − E[X])²) ∧ 0 < k ⇒
```

```
      prob p {x | x ∈ Ω ∧ kVar[X] ≤ |X x - E[X]|} ≤ $\frac{1}{k^2}$
```

The Markov inequality states that for any real number $k > 0$,

$$P(|X| \geq k) \leq \frac{m}{k} \tag{5.2}$$

Its formalization in HOL is the following:

```
 ⊢ random_variable X p Borel ∧ integrable p X ∧ 0 < k ⇒
      prob p {x | x ∈ Ω ∧ k ≤ |X x|} ≤ $\frac{E[X]}{k}$
```

Instead of directly proving these inequalities, we provide a more general proof using measure theory and Lebesgue integrals in HOL that can be used for both and a number of similar inequalities. The probabilistic statement follows by considering a space of measure 1.

Theorem 5.1. *Let $(S, \mathcal{S}, \mu)$ be a measure space, and let f be a measurable function defined on S. Then for any nonnegative function g, nondecreasing on the range of f,*

$$\mu(\{x \in S \,:\, f(x) \geq t\}) \leq \frac{1}{g(t)} \int_S g \circ f \, d\mu \,.$$

```
⊢ ∀m f g t.
      (let A = {x | x ∈ m_space m ∧ t ≤ f x} in
         measure_space m ∧
         f ∈ measurable (m_space m,measurable_sets m) Borel ∧
         (∀x. 0 ≤ g x) ∧ (∀x y. x ≤ y ⇒ g x ≤ g y) ∧
         integrable m (λx. g (f x)) ⇒
         measure m A ≤ (1 / (g t)) * integral m (λx. g (f x)))
```

The Chebyshev inequality is derived by letting $t = k\sigma$, $f = |X - m|$ and g defined as $g(t) = t^2$ if $t \geq 0$ and 0 otherwise. The Markov inequality is derived by letting $t = k$, $f = |X|$ and g defined as $g(t) = t$ if $t \geq 0$ and 0 otherwise.

Proof. Let $A = \{x \in S \,|\, t \leq f(x)\}$ and I_A be the indicator function of A. From the definition of A, $\forall x \; 0 \leq g(t)I_A(x)$ and $\forall x \in A, \; t \leq f(x)$. Since g is non-decreasing, $\forall x \; g(t)I_A(x) \leq g(f(x))I_A(x) \leq g(f(x))$. As a result, $\forall x \; g(t)I_A(x) \leq g(f(x))$. A is measurable because f is $(\mathcal{S}, \mathcal{B}(\overline{\mathbb{R}}))$ measurable. Using the monotonicity of the integral, verified in Chapter 3, $\int_S g(t)I_A(x)d\mu \leq \int_S g(f(x))d\mu$. Finally from the linearity of the integral $g(t)\mu(A) \leq \int_S g \circ f d\mu$. □

5.1.2 Weak Law of Large Numbers (WLLN)

The WLLN [49] states that the average of a large number of independent measurements of a random quantity converges in probability towards the theoretical average of that quantity. Interpreting this result, the WLLN states that for a sufficiently large sample, there will be a very high probability that the average will be close to the expected value. This law is used in a multitude of fields. It is used, for instance, to prove the AEP which is a fundamental concept in the field of information theory.

Theorem 5.2. *Let $X_1, X_2, ...$ be an infinite sequence of independent, identically distributed random variables with finite expected value $E[X_1] = E[X_2] = ... = m$ and let $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ then for any $\varepsilon > 0$,*

$$\lim_{n\to\infty} P(|\overline{X} - m| < \varepsilon) = 1 \tag{5.3}$$

$\vdash$ `prob_space p` $\wedge\ 0 < \varepsilon\ \wedge$

`(`$\forall$`i j. i`$\neq$`j` $\Rightarrow$ `uncorrelated p (`X_i`) (`X_j`))` $\wedge$

`(`$\forall$`i. E[`X_i`] = m)` $\wedge$ `(`$\forall$`i. Var[`X_i`] = v)` $\Rightarrow$

`lim (`λ`n. prob p {x | x` $\in \Omega\ \wedge$ `|`$\frac{1}{n}\sum_{i=1}^{n} X_i$ `x - m| <` ε`}) = 1`

Proof. Using the linearity property of the Lebesgue integral as well as the properties of the variance, we prove that $E[\overline{X}] = \frac{1}{n}\sum_{i=1}^{n} m = m$ and $Var(\overline{X}) = \frac{\sigma^2}{n}$. Applying the Chebyshev inequality to $\overline{X}$, we get $P(|\overline{X} - m| \geq \varepsilon) \leq \frac{\sigma^2}{n\varepsilon^2}$. Equivalently, $1 - \frac{\sigma^2}{n\varepsilon^2} \leq P(|\overline{X} - m| < \varepsilon) \leq 1$. It then follows that $\lim_{n\to\infty} P(|\overline{X} - m| < \varepsilon) = 1$. □

5.1.3 Asymptotic Equipartition Property (AEP)

The Asymptotic Equipartition Property (AEP) [14] is the information theory analog of the Weak Law of Large Numbers. It states that for a stochastic source X, if its time series $X_1, X_2, \ldots$ is a sequence of *iid* random variables with entropy $H(X)$, then $-\frac{1}{n}log(p(X_1, \ldots, X_n))$ converges in probability to $H(X)$.

Theorem 5.3. *(AEP): if $X_1, X_2, \ldots$ are iid then*

$$-\frac{1}{n}log(p(X_1, \ldots, X_n)) \longrightarrow H(X) \text{ in probability}$$

We formally verify the AEP using the WLLN result proved in the previous section as well as the various properties of joint probability distributions, independence of random variables and the log operator proved in Chapter 3.

```
⊢ prob_space p ∧ 0 < ε ∧
  (∀i j. i≠j ⇒ independent_rv p ($X_i$) ($X_j$)) ∧
  (∀i. E[$X_i$] = m) ∧ (∀i. Var[$X_i$] = v) ∧ (∀i. H[$X_i$] = H[X]) ⇒
  lim (λn. prob p {x | x∈ s ∧ |$-\frac{1}{n}log(\prod_{i=1}^{n} p_{X_i}\{x\})$ - H[X]| < ε}) = 1
```

Proof. Let $X_1, X_2, \ldots$ be *iid* random variables and let $Y_i = -log(p_{X_i})$. Then $Y_1, Y_2, \ldots$ are *iid* random variables and $\forall i$, $E[Y_i] = H(X)$. Using the Weak Law of Large Numbers, we have:

$$\lim_{n\to\infty} P(|\frac{1}{n}\sum_{i=1}^{n} Y_i - H(X)| < \varepsilon) = 1$$

Furthermore,

$$\frac{1}{n}\sum_{i=1}^{n} Y_i = \frac{1}{n}\sum_{i=1}^{n} -log(p(X_i)) = -\frac{1}{n}log(\prod_{i=1}^{n} p(X_i))$$

And since $X_1, \ldots, X_n$ are mutually independent,

$$-\frac{1}{n}log(\prod_{i=1}^{n} p(X_i)) = -\frac{1}{n}log(p(X_1 \ldots X_n))$$

Consequently,

$$\lim_{n\to\infty} P(| - \frac{1}{n}log(p(X_1 \ldots X_n)) - H(X)| < \varepsilon) = 1 \tag{5.4}$$

□

5.1.4 Typical Set

A consequence of the AEP is the fact that the set of observed sequences $(x_1, \ldots, x_n)$ for which joint probabilities $p(x_1, x_2, \ldots, x_n)$ are close to $2^{-nH(X)}$ has a total probability equal to 1. This set is called the *typical set* and such sequences are called the *typical sequences*. In other words, out of all possible sequences, only a small number of sequences will actually be observed and those sequences are nearly equally probable. The AEP guarantees that any property that is proved for the typical sequences will then be true with high probability and will determine the average behavior of a large sample.

Definition 5.1. *The typical set A^n_ε with respect to $p(x)$ is the set of sequences $(x_1, \ldots, x_n)$ satisfying:*

$$2^{-n(H(X)+\varepsilon)} \leq p(x_1, \ldots, x_n) \leq 2^{-n(H(X)-\varepsilon)}. \tag{5.5}$$

The typical set has the following properties

Theorem 5.4. *if* $(x_1, \ldots, x_n) \in A_\varepsilon^n$ *then*

$$H(X) - \varepsilon \leq -\frac{1}{n} log(p(x_1, \ldots, x_n)) \leq H(X) + \varepsilon \tag{5.6}$$

This theorem is a direct consequence of Definition 5.1.

Theorem 5.5. $\forall \varepsilon > 0, \exists N, \forall n \geq N, \ p(A_\varepsilon^n) > 1 - \varepsilon.$

The proof of this theorem is derived from the formally verified AEP. The next two theorems give upper and lower bounds for the number of typical sequences $|A_\varepsilon^n|$.

Theorem 5.6. $|A_\varepsilon^n| \leq 2^{n(H(X)+\varepsilon)}$.

Proof. Let $\underline{x} = (x_1, \ldots, x_n)$, then $\sum_{\underline{x} \in A_\varepsilon^n} p(\underline{x}) \leq 1$. From Equation 5.5, $\forall \underline{x} \in A_\varepsilon^n$, $2^{-n(H(X)+\varepsilon)} \leq p(\underline{x})$. Hence $\sum_{\underline{x} \in A_\varepsilon^n} 2^{-n(H(X)+\varepsilon)} \leq \sum_{\underline{x} \in A_\varepsilon^n} p(\underline{x}) \leq 1$. Consequently, $2^{-n(H(X)+\varepsilon)}|A_\varepsilon^n| \leq 1$ proving the theorem. □

Theorem 5.7. $\forall \varepsilon > 0, \exists N. \forall n \geq N, \ (1 - \varepsilon)2^{n(H(X)-\varepsilon)} \leq |A_\varepsilon^n|$.

Proof. Let $\underline{x} = (x_1, \ldots, x_n)$. From Theorem 5.5, $\exists N. \forall n \geq N, \ 1 - \varepsilon < \sum_{\underline{x} \in A_\varepsilon^n} p(\underline{x})$. From Equation 5.5, $\forall \underline{x} \in A_\varepsilon^n, \ p(\underline{x}) \leq 2^{-n(H(X)-\varepsilon)}$. Hence, $\exists N. \forall n \geq N, \ 1 - \varepsilon < \sum_{\underline{x} \in A_\varepsilon^n} p(\underline{x}) \leq \sum_{\underline{x} \in A_\varepsilon^n} 2^{-n(H(X)-\varepsilon)}$. Consequently, $\exists N. \forall n \geq N, \ 1-\varepsilon < 2^{-n(H(X)-\varepsilon)}|A_\varepsilon^n|$ proving the theorem. □

5.1.5 Data Compression Limit

The main idea behind the proof of the source coding theorem is that the average codeword length for all sequences is close to the average codeword length considering only the typical sequences. This is true because according to the typical set properties above, for a sufficiently large n, the typical set has a total probability close to 1. In other words, for any $\varepsilon > 0$, and sufficiently large n, the probability of observing a non-typical sequence is less than ε. Furthermore, the number of typical sequences is smaller than $2^{n(H(X)+\varepsilon)}$ and hence no more than $n(H(X)+\varepsilon)+1$ bits are needed to represent all typical sequences.

If we denote by Y the random variable defined over all the possible sequences and returns the corresponding codeword length. The expectation of the Y is equal to the average codeword length $\overline{L}$. Using the properties of the typical set we can prove that

$$\overline{L} \leq n(H(X)+\varepsilon^{'}) \tag{5.7}$$

where $\varepsilon^{'} = \varepsilon + \frac{1}{n}$.

Consequently, for any $\varepsilon > 0$ and n sufficiently large, the code rate $\frac{\overline{L}}{n}$ can be made as close as needed to the entropy $H(X)$ while maintaining a probability of error of the encoder that is bounded by ε.

5.1.6 Discussions

In this application, we made use of the framework presented in previous chapters for formally verifying the limit of data compression. This has the advantage of providing an exact mechanical proof of the result, similar to the one obtained through paper-and-pencil analysis, compared to less accurate results given by other computer based simulation approaches. Compared to the paper-and-pencil based analytical method, the correctness of the result is guaranteed by the soundness of the theorem prover. Furthermore, the formal proof serves as a way to improve the formal specification of the problem, by focusing only on the necessary assumptions to prove the result and ignoring the unnecessary ones.

To the best of our knowledge, this was the first time the AEP has been formally verified. This is also the first formal proof of Chebychev and Markov inequalities that uses measure theory and Lebesgue integration and that can be applied to various similar inequalities. The formalization of Shannon entropy, in Chapter 4, allowed us to define the typical set and prove its properties. These proofs required approximately 3 weeks of human effort and around 1000 lines of HOL code. The upside is that these results can be reused in several other engineering applications.

Our formalization has influenced the work of Affeldt et al [3], in which they proved the Shannon's theorems in the Coq proof assistants. In that work, the authors provided a simplified formalization of the concepts involved instead of generalized definitions that can be used in other applications, like we did in this thesis.

5.2 Anonymity-based Single MIX

In communication networks, privacy requires not only confidentiality of the information but also hiding the identities of the communicating parties. Several anonymous networks have been proposed to ensure anonymous communication, e.g. Onion Routing [65], Tor [20], Crowds [54], Mixminion [16], etc. Most of the proposed solutions are based on Chaum's original idea of a threshold mix [11]. Mixes are relay nodes that collect packets from different users, shuffle them then forward them to their destinations in such a way that an external eavesdropper cannot link an outgoing packet to its corresponding sender.

In this section, we use our formalization to reason about an anonymity-based single MIX, designed to hide the communication links between a set of senders and a set of receivers. We model a single MIX as a communication node connecting m senders $(s_1, \dots, s_m)$ to n receivers $(r_1, \dots, r_n)$. The single MIX is determined by its inputs (senders), outputs (receivers) and the transition probabilities. We can also add clauses in the specification to capture additional information about the MIX like structural symmetry. The following is the formalization of the single MIX given in Figure 5.3.

$\vdash$ `MIX_channel s m X Y =`

`X(s) = {0;1}` $\wedge$ `Y(s) = {0;1;2;3}` $\wedge$

`(`$p_{Y|X}$ `{0}{0} =` $\frac{1}{2}$`)` $\wedge$ `(`$p_{Y|X}$ `{1}{0} =` $\frac{1}{2}$`)` $\wedge$ `(`$p_{Y|X}$ `{2}{1} = 1)`

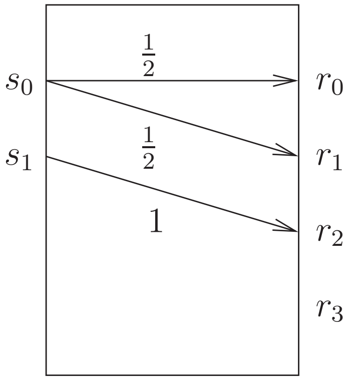

Figure 5.3: Single MIX Example

Zhu and Bettati [68] used the single MIX to model an anonymity-based covert-channel where a sender is trying to covertly send messages through the MIX. They used the channel capacity as a measure of the maximum information that can be leaked through the MIX and can be used as a measure of the quality of anonymity of the network. A communication between a sender s_i and a receiver r_j is denoted by $[s_i, r_j]$. The term $p([s_u, r_v]_s|[s_i, r_j]_a)$ represents the probability that the communication $[s_u, r_v]$ is suspected given that $[s_i, r_j]$ is actually taking place. This model describes attacks on sender-receiver anonymity. The input symbols of the covert-channel are the actual sender-receiver pairs $[s, r]_a$ and the output symbols are the suspected pairs $[s, r]_s$. In this case, $p([s, r]_s|[s, r]_a)$ represents the result of the anonymity attack. We consider the case where an attacker can establish a covert-channel by having 1 sender s_1 communicate with any combination of j receivers. The same reasoning can be applied to multiple senders. The authors claim the following result [68]:

For a single sender s_1 on a single mix, the maximum covert-channel capacity is achieved when s_1 can communicate to all receivers.

We initially tried to formally verify this result, using the foundational results presented in the previous chapters, but we found a counter-example for an assumption upon which the paper-and-pencil proof of the above result is based. The erroneous assumption states that the maximum of the mutual information is achieved when all input symbols have non-zero probabilities regardless of the transition probabilities (the results of the anonymity attack). We are able to prove in HOL that it is not necessary for the sender s_1 to communicate with all receivers to achieve capacity.

First, we provide a higher-logic-formalization of the channel capacity which is defined as the maximum, over all input distributions, of the mutual information between the input and the output of the channel. We formalize it in HOL using the formalization of mutual information from Chapter 4 and the Hilbert-choice operator; i.e., if it exists, the capacity is some c such that $c = I_m(X;Y)$ for some probability distribution m and for any input distribution p, $I_p(X;Y) \leq c$.

$\vdash$ `capacity s X Y = @c. (`$\exists$`m. c = `$I_m(X;Y)$`) `$\wedge$` (`$\forall$`m. `$I_m(X;Y)$` `$\leq$` c)`

Next, consider the covert-channel depicted in Figure 5.4. To simplify the notation, let $x_i = [s_1, r_i]_a$ and $y_i = [s_1, r_i]_s$. This covert-channel is formalized in HOL as

$\vdash$ `MIX_channel_1 s m X Y =`

  `(X(s) = {0;1;2})` $\wedge$ `(Y(s) = {0;1;2})` $\wedge$

  `(`p_X`{0} = `p_X`{2})` $\wedge$

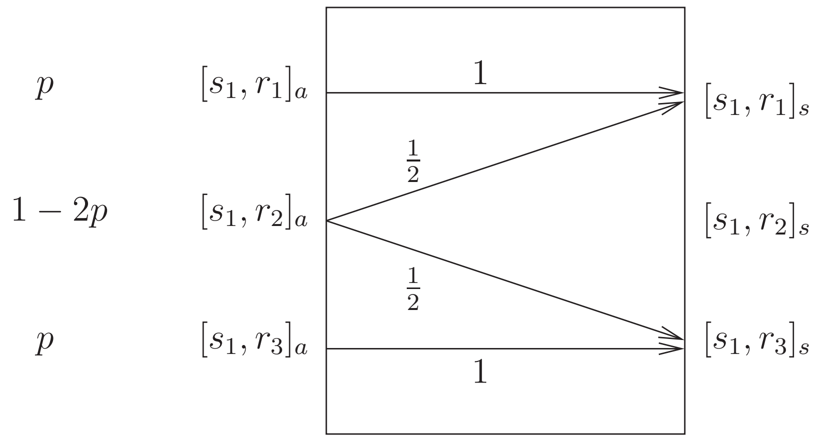

Figure 5.4: Counter-Example for [68]

$(p_{Y|X}\ \{0\}\{0\} = 1) \wedge (p_{Y|X}\ \{0\}\{1\} = \frac{1}{2}) \wedge (p_{Y|X}\ \{0\}\{2\} = 0) \wedge$

$(p_{Y|X}\ \{1\}\{0\} = 0) \wedge (p_{Y|X}\ \{1\}\{1\} = 0) \wedge (p_{Y|X}\ \{1\}\{2\} = 0) \wedge$

$(p_{Y|X}\ \{2\}\{0\} = 0) \wedge (p_{Y|X}\ \{2\}\{1\} = \frac{1}{2}) \wedge (p_{Y|X}\ \{2\}\{2\} = 1)$

We prove that its mutual information is equal to $2p$.

```
⊢ ∀X Y s. MIX_channel_1 s m X Y ⇒
        I(X;Y) = 2 * p X {0}
```

We also prove that the capacity is equal to 1 and corresponds to $p = \frac{1}{2}$. This means that the input distribution that achieves the channel capacity is $[p\{x_0\} = \frac{1}{2},\ p\{x_1\} = 0,\ p\{x_2\} = \frac{1}{2}]$. Hence, we prove that the sender s_1 does not need to communicate with the receiver r_2 and still achieve maximum capacity, contradicting the result in [68]. Notice that with $p = \frac{1}{2}$, $I(X;Y) = H(X) = 1$ which implies that the degree of information leakage $D = 0$. So for this covert-channel, the maximum capacity corresponds to perfect identification.

5.2.1 Discussions

Unlike the paper-and-pencil based analysis, a machine-assisted analysis of quantitative information flow using theorem proving guarantees the accuracy of the results. In fact, the soundness of theorem proving inherently ensures that only valid formulas are provable. The requirement that every single step of the proof needs to be derived from axioms or previous theorems using inference rules, allows us to find missing assumptions and even sometimes wrong statements as was the case in this single MIX application. We were able to detect the problem with the reasoning described in the above sections and confirm the result using our formalization in HOL. In this specific case, we detected the problem when trying to prove the erroneous assumption stating that the channel capacity is achieved when all input symbols have non-zero probabilities. This result contradicts Theorem 4.5.1 of [22], which inspired us to come up with the counter-example. Our analysis has also been confirmed by Prof. Gallager from MIT, the author of the much cited book *Information Theory and Reliable Communication* [22].

To the best of our knowledge, this is the first time the properties of mixes have been analyzed using theorem proving. This is obviously not a large application and can be extended to reason about MIX networks and other anonymity networks in general. However this serves as an example to illustrate the usefulness of the framework presented in this thesis. Thanks to the rich formalization we provided in Chapters 3 and 4, we were able to analyze the MIX of this application within one week of human effort and using around 500 lines of HOL code.

5.3 One-Time Pad

The one-time pad is a simple yet solid encryption system that provides, if used correctly, an unbreakable security. The encryption is performed by modular addition of every character of the plaintext with a character from a secret random key of at least the same length as the original message. If the key is truly random and never reused in whole or in part, then it can be proven that the one-time pad encryption provides a perfect security. We formally prove this property within the HOL4 theorem prover using the higher-order-logic framework proposed in this thesis.

The one-time pad encryption technique takes its name from the paper pads that have been historically used to distribute the keys, making it easy to simply pull the top sheet of the pad and destroy it after use. An example of a Russian one-time pad that was captured by MI5 is depicted in Figure 5.5.

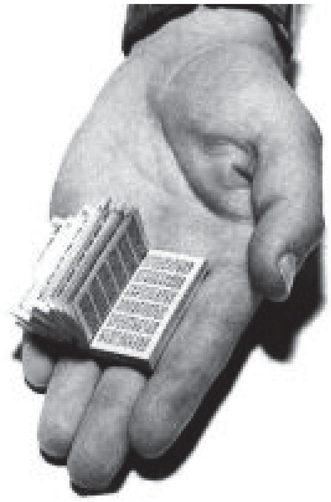

Figure 5.5: A Russian One-time pad, captured by MI5 [53]

The one-time pad has been extensively used to secure the communications of various international intelligence agencies and was used for instance in the Washington/Moscow hotline to provide perfectly secure communication between the White

House and the Kremlin and without disclosing any other secret cryptographic technology.

The main challenges for this encryption technique are the generation of truly random keys and their distribution to both sender and receiver. This sometimes makes the technique impractical and limits the types of its applications to the cases where, for example, absolute security is a real must, regardless of the costs. Still, the one-time pad is available as a backup encryption option if other theoretically less secure but more practical encryption systems are unavailable for reasons of war or attacks. The one-time pad encryption is also very important in situations, where both sender and receiver need to do all the work by hand without the use of a computer, whether because one is not available or to avoid possible vulnerabilities of a standard computer.

The structure of a typical one-time pad encryption system is depicted in Figure 5.6. The plaintext is first encoded into digits or bits then fed to the encryption block which performs a modular addition (modulo 10) to produce a cipher text. The latter is transmitted to the receiver side which performs the inverse operations to recover the original message.

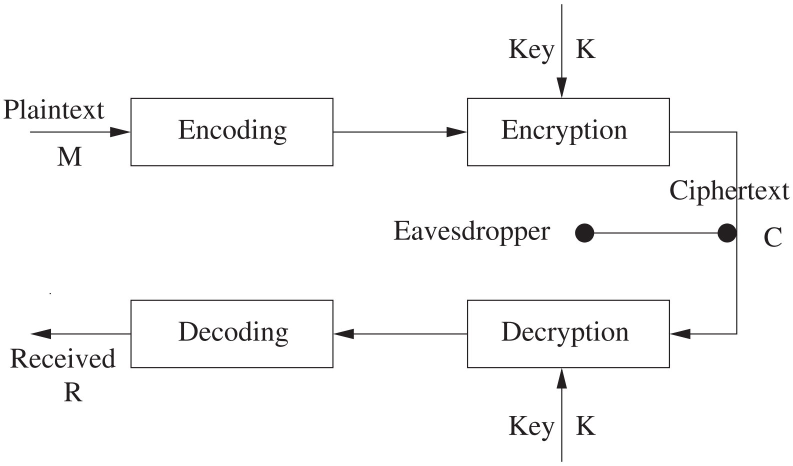

Figure 5.6: One-Time Pad Encryption

	0	1	2	3	4	5	6	7	8	9
	A	T		O	N	E		S	I	R
2	B	C	D	F	G	H	J	K	L	M
6	P	Q	U	V	W	X	Y	Z	.	/

Table 5.1: Straddling Checkerboard Example

5.3.1 Encoding - Decoding

We use a straddling checkerboard to convert the alphabetic plaintext into digits. With this conversion scheme, the more frequent letters in a language are encoded with a lower number of digits, leading to a compressed output and, hence, shorter messages to be transmitted. Besides, a straddling checkerboard allows to achieve a simple form of information diffusion, or in other words, it reduces the redundancy in the statistics of the plaintext. An example checkerboard for the English language can be found in Table 5.1. We formalize the straddling checkerboard as the function `checkerboard` of the HOL type:

```
⊢ checkerboard: char -> num
```

We present the definition of `checkerboard` associated with Table 5.1 for the first-row letters as well as `P` and `/`.

```
⊢ (checkerboad #’’A’’ = 0)  ∧
  (checkerboad #’’T’’ = 1)  ∧
  (checkerboad #’’O’’ = 3)  ∧
  (checkerboad #’’N’’ = 4)  ∧
  (checkerboad #’’E’’ = 5)  ∧
  (checkerboad #’’S’’ = 7)  ∧
  (checkerboad #’’I’’ = 8)  ∧
```

```
(checkerboad #’’R’’ = 9)  ∧
(checkerboad #’’P’’ = 60) ∧
(checkerboad #’’/’’ = 69)
```

Using the above definition of the straddling checkerboard, we formalize the encoding and decoding blocks as `encode` and `decode` functions, respectively. The encoder takes as input a string representing the alphabetic plaintext which it decomposes into a list of characters, each of which is processed through the checkerboard, and returns a list of digits. The decoder performs the inverse operations to convert a list of digits back to a string. The functions `encode` and `decode` have the following HOL types:

```
⊢ encode: string → num list
⊢ decode: num list → string
```

5.3.2 Encryption - Decryption

The ecryption and decryption blocks are formalized as two functions, `encrypt` and `decrypt`, taking as input a pair of same length lists of digits and returning a list of digits.

```
⊢ encrypt:(num list,num list) → num list
⊢ decrypt:(num list,num list) → num list
```

The encryption is performed by a *modulo*10 addition, digit by digit, of the list representing the encoded message and the list of digits representing the one-time pad key. The result of this operation is a ciphertext which is also represented by a list of digits. On the receiver side, the ciphertext is decrypted by subtracting, *modulo*10, the key from ciphertext, resulting into a list of numbers that represent the original message. In the case where the plaintext is encoded into bits instead of digits, both encryption and decryption are performed by a simple XOR operation. We formalize `encrypt` in higher-order logic, recursively. h_1 and h_2 represent the first elements or heads of the lists and t_1 and t_2 their tails. The :: operator is the list constructor.

```
⊢  encrypt ([],[]) = [] ∧
    ∀t1 t2 h1 h2.
   encrypt (h1::t1,h2::t2) =
       (h1+h2) MOD 10::encrypt (t1,t2)
```

Similarly, we formalize the decryption block as follows.

```
⊢  decrypt ([],[]) = [] ∧
    ∀t1 t2 h1 h2.
   decrypt (h1::t1,h2::t2) =
       (h1-h2) MOD 10::decrypt (t1,t2)
```

Finally, let m be the original message (plaintext), k be the one-time pad key and r be the received message after decryption and decoding. The one-time pad (OTP) encryption is then formalized in HOL using the following predicate.

```
⊢ ∀ m k r.
    OTP m k r ⇔
r = decode(decrypt(encrypt(encode m,k),k))
```

As a reassuring property, we prove in HOL that the one-time pad as designed and formalized above, ensures that the received message is equal to the original message.

```
⊢ ∀ m k r. OTP m k r ⇒ (r = m)
```

5.3.3 Perfect Security

We formally verify that the one-time pad provides perfect security by proving that the information leakage degree, formalized in Chapter 4 Section 4.3, is equal to one.

Let M, C and K denote the random variables representing the plaintext, ciphertext and keys, respectively. Hence, K is uniformly distributed and is independent of M, which allows us to prove that

$$\vdash \quad \forall\ \mathtt{m} \in \mathcal{M},\ \mathtt{c} \in \mathcal{C}.$$
$$\mathtt{P(M = m \mid C = c) = P(M = m)}$$

This follows from the following lemmas, which we prove using the properties we formally proved in Chapter 4 about probability distributions.

$\vdash \quad \mathtt{P(M = m \mid C = c) = P(M = m,\ C = c)\ /\ P(C = c)}$

$\vdash \quad \mathtt{P(M = m,\ C = c) = P(M = m,\ K = m} \oplus \mathtt{c)}$

$\vdash \quad \mathtt{P(M = m,\ K = m} \oplus \mathtt{c) = P(M = m)\ P(K = m} \oplus \mathtt{c)}$

$\vdash \quad \mathtt{P(K = m} \oplus \mathtt{c) =}\ 2^{-n}$

$\vdash \quad \mathtt{P(C = c) =}\ 2^{-n}$

Next, we prove that the conditional entropy of M given C is equal to the entropy of M and that the mutual information $I(M;C)$ is equal to zero.

$\vdash \quad \mathtt{H(M|C) = H(M)}$

$\vdash \quad \mathtt{I(M;C) = 0}$

Finally, it follows that the information leakage degree is equal to 1, meaning that the one-time pad encryption is information-theoretically secure and there is no leakage of information about the secret input (plaintext) to a possible eavesdropper.

$\vdash \quad \mathtt{D(M,C) = 1}$

5.3.4 Discussions

In this application, we were able to formally prove the perfect security property of the OTP encryption system thanks to the various properties of probability distributions presented in Section 4.1 of Chapter 4 as well as the properties of the Shannon entropy and mutual information from Section 4.2 of the same Chapter. Theorem proving allows to provide a generic result that does not depend on which message has been encrypted, unlike the kind of results produced by computer simulation. If fact, simulation can be used to detect the presence of bugs but is not useful to guarantee their absence. The formalization of the different components of the OTP as well as the proof of its security property required around two weeks of human effort and around 800 lines of code.

5.4 Summary and Discussions

In this chapter, we have verified some classical results of probability theory, namely, the Shebyshev and Markov inequalities and the Weak Law of Large Numbers. We used these results to formally verify the Asymptotic Equipartition Property, an important property used in the proofs of numerous information-theoretic theorems. We use the AEP to verify the properties of a typical encoder that is used in the formal proof of the Shannon source coding theorem. We have also presented two example applications of the use of quantitative analysis within a theorem prover to analyse the properties of an anonymity-based MIX channel as well as the properties of the one-time pad encryption. In the first example, we were able to detect a problem and come up with a counter-example to a result that was reported in a prominent paper [68]. In the second example, we were able to formalize the encryption system, verify its functionality as well as prove a generic result about its perfect security. The soundness and the deduction style of the theorem prover guarantee the validity of the analysis when deriving these proofs. Besides, the results of this type of analysis are generic and valid for any instance of the system. We argue that these benefits are even more significant when dealing with larger and more complex systems as is the case for nowadays parallel and distributed systems.

These applications illustrate how our formalization of information theory and the different underlying theories of measure, Lebesgue integration and probability, can be used to reason about a multitude of engineering applications. Conducting the analysis within the sound core of a theorem prover helped to add more trust to the proved results. It allowed to detect a bug in the paper-and-pencil analysis of the MIX channel example. While the formalization of the different theories required more than 20,000 lines of code, the applications of this chapter have only required around 800 lines of code, on average. Still, a considerable effort by the user is still needed to conduct the analysis.

Chapter 6

Conclusions and Future Work

6.1 Conclusions

In this thesis, we have proposed to conduct the information-theoretic analysis of systems and protocols using a higher-order-logic theorem prover. To do so, we proposed a formal framework of all the tools needed to describe the system under consideration and its desired properties in a formal language that can be used within the theorem prover and make it possible to formally analyze these properties. Compared to the standard techniques of computer simulation and paper-and-pencil analysis, our approach allows to exploit the soundness of theorem proving to deliver more accurate and trusted results. It also allows to provide generic results instead of proving the properties for specific instances of the system.

The framework we provided consists in a higher-order-logic formalization of the main concepts of information theory along with the formalization of the underlying theories of measure, Lebesgue integration and probability. We formalized the most commonly used measures of information starting with the KL divergence which we used to define the other measures like the Shannon entropy and mutual information.

This provides the tools necessary to perform the quantitative analysis of information for various applications like the evaluation of performance of various anonymity and security protocols.

We provided generalized formalizations of measure and Lebesgue integration which are based on the extended-real numbers and the Borel sigma algebras. Using this, we formalized the main concepts of probability and verified a number of classical results. The way these theories have been formalized, allows to work with discrete and continuous random variables in a unified framework where the properties are valid for both cases. We then provided simplifications of the theorems in the cases of finite or countable spaces.

While we built on previous research for the different theories, to the best of our knowledge, our work provides the most generalized formalizations as has been discussed in the previous chapters. Furthermore, we introduced, in this thesis, two novel measures of information leakage that can be used, for instance, in the evaluation of anonymity and privacy protocols. Our work has been accepted to be part of the official release of the HOL4 theorem prover, but it can also be adopted for any other higher-order-logic theorem prover.

Finally, we illustrated the usefulness of our approach and formal framework by tackling several applications in the areas of communication, anonymity and privacy. We proved the properties of a typical encoder used in the proof of the Shannon source coding theorem. We also analyzed an anonymity based single MIX that was proposed in the prominent paper of Zhu and Bettati [68]. We were able to find an error in the paper-and-pencil analysis presented in the paper thanks to the soundness of theorem proving. We offered a counter-example to the result proposed in [68] to confirm our analysis. Finally, we used our framework to formally prove the perfect security property of the one-time pad encryption. These applications highlight the feasibility and benefits of conducting the information-theoretic analysis using a higher-order-logic theorem prover. In fact, the added trust provided by the deduction style of theorem proving, is a crucial requirement when dealing with safety-critical applications. Another benefit of this approach is the generic results that are guaranteed to be valid for every single instance of the system. These benefits are even more significant for larger and more complex systems.

6.2 Future Work

Information theory has been an important basis for the analysis of a wide range of applications, especially in the fields of communication and cryptography. This thesis lays the ground to a promising approach for the information-theoretic analysis of safety-critical applications. Building on the formalization and verification results presented in this thesis, several extensions can be explored to further strengthen the proposed framework. Some future research directions are outlined below.

- While using the extended-real numbers made it possible to provide a more generalized formalization of measure and Lebesgue integration, it added some complexity to the proofs. It is possible to simplify these proofs by creating a simplification set `ext_ss` which contains the most used theorems of the extended-real numbers theory, especially the various properties of the operators. Creating tactics and decision procedures for this purpose is also an interesting extension that needs to be explored.

- The probability density function (PDF) is defined as the Radon Nikodym derivative of the cumulative distribution function with respect to the probability measure. Formalizing the PDF allows to analyze the class of systems that can be described by standard continuous distributions such as the normal distribution. The latter can be used to model a wide class of systems due to the central limit theorem. A formal proof of this theorem can also be an interesting extension to the formalization developed in this thesis. Finally, formalizing the PDF allows to write the statistical properties of random variables, like the expectation, in terms of the density function and hence can be useful in the evaluation of continuous systems.

- The definition of the Borel sigma algebra is based on the open sets which can be used to define the Lebesgue integral for functions ranging over Euclidian spaces. On the other hand, our formalization is general enough to handle functions of multiple variables. The integral of these functions is what is called the multiple integral. The formalization can be enriched by proving the properties of multiple integrals such as under which conditions it is possible to change the order of the integrals, which can be useful in many cases to simplify the computation of the multiple integral.

- Our formalization can also be enriched by porting some of the concepts that have been developed by Abbasi [1] to perform the reliability analysis of engineering systems. We believe that our formalization of probability is more suited to conduct this analysis since it allows to prove the general properties of the system regardless of which probability distribution is used. The unified framework for discrete and continuous random variables is another advantage of this porting. More importantly, our formalization does not require the independence of random variables and hence can be used to analyse a larger class of systems.

- Crowds [54] and Tor [20] are two of the main solutions proposed to provide anonymity in communication networks. The evaluation of the performance of these networks is an interesting application of the approach proposed in this thesis. The measures of information leakage proposed in this thesis can be used to quantify the leakage of information for different configuration of these networks. Conducting the analysis using theorem proving provides accurate analysis and generic results.

- Building on the data compression application, jointly typical sets and the notion of communication channel and codes can be formalized allowing to prove the Shannon channel coding theorem as well as other coding theorems. This will constitutes an important step to formalize digital communication systems and evaluate their performances.

- A mix network [59] consists of several interconnected stages or single mixes. the interconnection of the stages determines the network topology. The single mix application of this thesis can be extended by formalizing mix networks as well as the basic attacks and adversary models in the networks. This will allow to reason about several real-world applications like electronic voting and anonymous email and telecommunications.

Bibliography

[1] N. Abbasi. *Formal Reliability Analysis using Higher-Order Logic Theorem Proving*. PhD thesis, Concordia University, Montreal, Quebec, Canada, 2012.

[2] J. R. Abrial. Faultless Systems: Yes We Can! *IEEE Computer Journal*, 42(9):30–36, 2009.

[3] R. Affeldt and M. Hagiwara. Formalization of Shannon's Theorems in SSReflect-Coq. In *Interactive Theorem Proving*, volume 7406 of *LNCS*, pages 233–249. Springer, 2012.

[4] C. Baier, B. Haverkort, H. Hermanns, and J. P. Katoen. Model Checking Algorithms for Continuous time Markov Chains. *IEEE Transactions on Software Engineering*, 29(4):524–541, 2003.

[5] C. Baier and J. P. Katoen. *Principles of Model Checking*. MIT Press, 2008.

[6] S. K. Berberian. *Fundamentals of Real Analysis*. Springer, 1998.

[7] Y. Bertot and P. Casteran. *Coq'Art: The Calculus of Inductive Constructions*. Springer, 2004.

[8] J. Bialas. The σ-Additive Measure Theory. *Journal of Formalized Mathematics*, 2(2):263–270, 1991.

[9] V. I. Bogachev. *Measure Theory*. Springer, 2006.

[10] K. Chatzikokolakis, C. Palamidessi, and P. Panangaden. Anonymity Protocols as Noisy Channels. In *Trustworthy Global Computing*, volume 4661 of *LNCS*, pages 281–300. Springer, 2007.

[11] D. L. Chaum. Untraceable Electronic Mail, Return Addresses, and Digital Pseudonyms. *Communications of the ACM*, 24(2):84–90, February 1981.

[12] A. Church. A Formulation of the Simple Theory of Types. *Journal of Symbolic Logic*, 5:56–68, 1940.

[13] A. R. Coble. *Anonymity, Information, and Machine-Assisted Proof*. PhD thesis, University of Cambridge, Cambridge, UK, 2010.

[14] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley-Interscience, 1991.

[15] Crypto Museum. One-Time Pad (OTP), the Unbreakable Code. http://www.cryptomuseum.com/crypto/otp.htm, 2010.

[16] G. Danezis, R. Dingledine, D. Hopwood, and N. Mathewson. Mixminion: Design of a Type III Anonymous Remailer Protocol. In *IEEE Symposium on Security and Privacy*, pages 2–15, 2003.

[17] L. de Alfaro. *Formal Verification of Probabilistic Systems*. PhD Thesis, Stanford University, Stanford, California, USA, 1997.

[18] Y. Deng, J. Pang, and P. Wu. Measuring Anonymity with Relative Entropy. In *Formal Aspects in Security and Trust*, volume 4691 of *LNCS*, pages 65–79. Springer, 2007.

[19] C. Diaz, S. Seys, J. Claessens, and B. Preneel. Towards Measuring Anonymity. In *Privacy Enhancing Technologies*, volume 2482 of *LNCS*, pages 54–68. Springer, 2003.

[20] R. Dingledine, N. Mathewson, and P. Syverson. Tor: The Second-Generation Onion Router. In *USENIX Security Symposium*, pages 303–320, 2004.

[21] A. A. Fraenkel, Y. Bar-Hillel, and A. Levy. *Foundations of Set Theory*. North Holland, 1973.

[22] Robert G. Gallager. *Information Theory and Reliable Communication*. John Wiley & Sons, Inc., 1968.

[23] R. R. Goldberg. *Methods of Real Analysis*. Wiley, 1976.

[24] M. J. C. Gordon and T. F. Melham. *Introduction to HOL: A Theorem Proving Environment for Higher-Order Logic*. Cambridge University Press, 1993.

[25] R. M Gray. *Entropy and Information Theory*. Springer-Verlag, 1990.

[26] P. R. Halmos. The foundations of probability. *The American Mathematical Monthly*, 51(9):493–510, 1944.

[27] J. Harrison. Formalized Mathematics. Technical Report 36, Turku Centre for Computer Science, Finland, 1996.

[28] J. Harrison. *Theorem Proving with the Real Numbers*. Springer, 1998.

[29] J. Harrison. A HOL Theory of Euclidean Space. In *Theorem Proving in Higher Order Logics*, volume 3603 of *LNCS*, pages 114–129. Springer, 2005.

[30] J. Harrison. *Handbook of Practical Logic and Automated Reasoning*. Cambridge University Press, 2009.

[31] O. Hasan. *Formal Probabilistic Analysis using Theorem Proving*. PhD thesis, Concordia University, Montreal, Quebec, Canada, 2008.

[32] O. Hasan, N. Abbasi, B. Akbarpour, S. Tahar, and R. Akbarpour. Formal Reasoning about Expectation Properties for Continuous Random Variables. In *Formal Methods*, volume 5850 of *LNCS*, pages 435–450, 2009.

[33] O. Hasan and S. Tahar. Verification of Expectation Properties for Discrete Random Variables in HOL. In *Theorem Proving in Higher-Order Logics*, volume 4732 of *LNCS*, pages 119–134. Springer, 2007.

[34] O. Hasan and S. Tahar. Formal Verification of Tail Distribution Bounds in the HOL Theorem Prover. *Mathematical Methods in the Applied Sciences*, 32(4):480–504, March 2009.

[35] O. Hasan and S. Tahar. Performance Analysis and Functional Verification of the Stop-and-Wait Protocol in HOL. *Journal of Automated Reasoning*, 42(1):1–33, 2009.

[36] J. Hölzl and A. Heller. Three Chapters of Measure Theory in Isabelle/HOL. In *Interactive Theorem Proving*, volume 6898 of *LNCS*, pages 135–151. Springer, 2011.

[37] J. Hurd. *Formal Verifcation of Probabilistic Algorithms*. PhD thesis, University of Cambridge, Cambridge, UK, 2002.

[38] J. Hurd, A. McIver, and C. Morgan. Probabilistic Guarded Commands Mechanized in HOL. *Electronic Notes in Theoretical Computer Science*, 112:95–111, 2005.

[39] A. N. Kolmogorov. *Grundbegriffe der Wahrscheinlichkeitsrechnung.* Springer, 1933. English translation (1950): Foundations of the Theory of Probability. Chelsea Publishing Co.

[40] R. P. Kurshan. Formal Verification in a Commercial Setting. In *Design Automation Conference*, pages 258–262. ACM, 1997.

[41] M. Kwiatkowska, G. Norman, and D. Parker. Quantitative Analysis with the Probabilistic Model Checker PRISM. *Electronic Notes in Theoretical Computer Science*, 153(2):5–31, 2005.

[42] D. Lester. Topology in PVS: Continuous Mathematics with Applications. In *Workshop on Automated Formal Methods*, pages 11–20. ACM, 2007.

[43] P. Manolios M. Kaufmann and J. S. Moore. *Computer-Aided Reasoning: An Approach.* Kluwer Academic Publishers, 2000.

[44] P. Malacaria. Assessing Security Threats of Looping Constructs. *SIGPLAN Notes*, 42(1):225–235, 2007.

[45] R. Milner. A Theory of Type Polymorphism in Programming. *Journal of Computer and System Sciences*, 17:348–375, 1977.

[46] A. Nędzusiak. σ-fields and Probability. *Journal of Formalized Mathematics*, 1, 1989.

[47] Kananaskis-8 Release Notes on HOL 4. http://hol.sourceforge.net/kananaskis-8.release.html, 2012.

[48] S. Owre, J. M. Rushby, , and N. Shankar. PVS: A Prototype Verification System. In *Automated Deduction*, volume 607 of *LNCS*, pages 748–752. Springer, 1992.

[49] A. Papoulis. *Probability, Random Variables, and Stochastic Processes*. Mc-Graw Hill, 1984.

[50] D. Parker. *Implementation of Symbolic Model Checking for Probabilistic System*. PhD Thesis, University of Birmingham, Birmingham, UK, 2001.

[51] L. C. Paulson. *Isabelle: a Generic Theorem Prover*. Springer, 1994.

[52] L.C. Paulson. *ML for the Working Programmer*. Cambridge University Press, 1996.

[53] M. J. Ranum. One-Time Pad FAQ.
http://www.ranum.com/security/computer_security/papers/otp-faq/, 1995.

[54] M. K. Reiter and A. D. Rubin. Crowds: Anonymity for Web Transactions. *ACM Transactions on Information and System Security*, 1(1):66–92, 1998.

[55] S. Richter. Formalizing Integration Theory with an Application to Probabilistic Algorithms. In *Theorem Proving in Higher Order Logics*, volume 3223 of *LNCS*, pages 271–286. Springer, 2004.

[56] W. Rudin. *Real and Complex Analysis*. McGraw-Hill, 1987.

[57] J. Rutten, M. Kwaiatkowska, G. Normal, and D. Parker. *Mathematical Techniques for Analyzing Concurrent and Probabilisitc Systems*, volume 23 of *CRM Monograph Series*. American Mathematical Society, 2004.

[58] A. Sabelfeld and A. C. Myers. Language-Based Information-Flow Security. *IEEE Journal on Selected Areas in Communications*, 21(1):5–19, 2003.

[59] K. Sampigethaya and R. Poovendran. A Survey on Mix Networks and Their Secure Applications. *Proceedings of the IEEE*, 94(12):2142–2181, 2006.

[60] K. Sen, M. Viswanathan, and G. Agha. VESTA: A Statistical Model-Checker and Analyzer for Probabilistic Systems. In *Quantitative Evaluation of Systems*, pages 251–252. IEEE Computer Society, 2005.

[61] A. Serjantov and G. Danezis. Towards an Information Theoretic Metric for Anonymity. In *Privacy Enhancing Technologies*, volume 2482 of *LNCS*, pages 259–263. Springer, 2003.

[62] C. E. Shannon. A Mathematical Theory of Communication. *The Bell System Technical Journal*, 27(3):379–423, 1948.

[63] Y. Shidama, N. Endou, and P.N. Kawamoto. On the formalization of lebesgue integrals. *Studies in Logic, Grammar and Rhetoric*, 10(23):167–177, 2007.

[64] G. Smith. On the Foundations of Quantitative Information Flow. In *Foundations of Software Science and Computational Structures*, volume 5504 of *LNCS*, pages 288–302. Springer, 2009.

[65] P. F. Syverson, D. M. Goldschlag, and M. G. Reed. Anonymous Connections and Onion Routing. In *Symposium on Security and Privacy*, pages 44–54. IEEE Computer Society, 1997.

[66] HOL4 Theorem Prover. http://hol.sourceforge.net/, 2012.

[67] S. Wagon. *The Banach-Tarski Paradox*. Cambridge University Press, 1993.

[68] Ye Zhu and Riccardo Bettati. Information Leakage as a Model for Quality of Anonymity Networks. *IEEE Transactions on Parallel and Distributed Systems*, 20(4):540–552, 2009.

Biography

Education

- **Concordia University**: Montreal, Quebec, Canada

  Ph.D candidate, Electrical & Computer Engineering, (Sep. 09 - present)

- **Concordia University**: Montreal, Quebec, Canada

  M.A.Sc, Electrical & Computer Engineering, (Sep. 01 - Aug. 03)

- **Ecole Polytechnique de Tunisie**: La Marsa, Tunis, Tunisia

  B.Sc, Electrical & Computer Engineering, (Sep. 98 - Jun. 01)

- **Institut Préparatoire IPEST**: La Marsa, Tunis, Tunisia

  Diploma in Math and Physics Studies, (Sep. 96 - Sep. 98)

Awards

- Tunisia National Bursary for PhD study in Canada (2003-2006)

- Tunisia National Bursary for MASc study in Canada (2001-2003)

Work History

- **Concordia University**: Montreal, Quebec, Canada

  Research Assitant, Electrical & Computer Engineering (2009-2012)

- **enQuira, inc.**: Montreal, Quebec, Canada

  CEO and Founder, Local Search and e-Marketing (2006-2012)

- **McGill University - InterDigital, inc.**: Montreal, Quebec, Canada

  Research Assitant, Electrical Engineering (2003-2006)

- **Concordia University**: Montreal, Quebec, Canada

  Research Assitant, Electrical & Computer Engineering (2001-2003)

Publications

- **Journal Papers**

  - **Bio-Jr1** T. Mhamdi, O. Hasan, and S. Tahar. "Formalization of Measure Theory and Lebesgue Integration for Probabilistic Analysis in HOL", *ACM Transactions on Embedded Computing Systems*, Accepted in June 2011. [25 pages].

  - **Bio-Jr2** T. Mhamdi, O. Hasan, and S. Tahar. "Performance Evaluation of Anonymity and Security Protocols using Theorem Proving", *IEEE Transactions on Parallel and Distributed Systems*, Submitted in September 2012. [12 pages].

- **Refereed Conference Papers**

  - **Bio-Cf1** T. Mhamdi, O. Hasan, and S. Tahar. “Quantitative Analysis of Information Flow using Theorem Proving”, In *International Conference on Formal Engineering Methods*, volume 7635 of *LNCS*, pages 119–134. Springer, 2012.

  - **Bio-Cf2** T. Mhamdi, O. Hasan, and S. Tahar. “Formalization of Entropy Measures in HOL”, In *Interactive Theorem Proving*, volume 6898 of *LNCS*, pages 233–248. Springer, 2011.

  - **Bio-Cf3** T. Mhamdi, O. Hasan, and S. Tahar. “On the Formalization of the Lebesgue Integration Theory in HOL”, In *Interactive Theorem Proving*, volume 6172 of *LNCS*, pages 387–402. Springer, 2010.

  - **Bio-Cf4** T. Mhamdi and S. Tahar: “Providing Automated Verification in HOL using MDGs”, In *Automated Technology for Verification and Analysis*, volume 3299 of *LNCS*, pages 278-293. Springer, 2004.

  - **Bio-Cf5** T. Mhamdi and S. Tahar: Embedding Multiway Decision Graphs in HOL; B-Track, In *International Conference on Theorem Proving in Higher-Order Logics*, (TPHOLs’04), Park City, Utah, USA, September 2004, pp. 121-136

- **Technical Report**

  - **Bio-Tr1** T. Mhamdi, O. Hasan, and S. Tahar, “Formalization of Measure and Lebesgue Integration over Extended Reals in HOL”, Technical Report, Department of Electrical and Computer Engineering, Concordia University, January 2011. [120 Pages].

  - **Bio-Tr2** T. Mhamdi, O. Hasan, and S. Tahar. “On the Formalization of the Lebesgue Integration Theory in HOL”, Technical Report, Department of Electrical and Computer Engineering, Concordia University, January 2010. [25 Pages].

- **Bio-Tr3** T. Mhamdi, and S. Tahar: "Embedding Multiway Decision Graphs in HOL"; Technical Report, Concordia University, Department of Electrical andComputer Engineering, February 2004. [20 pages]