Formalization of Entropy Measures in HOL

Tarek Mhamdi, Osman Hasan, and Sofiène Tahar

ECE Department, Concordia University, Montreal, QC, Canada
{mhamdi,o_hasan,tahar}@ece.concordia.ca

Abstract. Information theory is widely used in a very broad class 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 and thus cannot be completely relied upon. To facilitate the formal reasoning about information theoretic aspects, this paper presents a rigorous higher-order logic formalization of some of the most widely used information theoretic principles. Building on fundamental formalizations of measure and Lebesgue integration theories for extended reals, we formalize the Radon-Nikodym derivative and prove some of its properties using the HOL theorem prover. This infrastructure is then used to formalize information theoretic fundamentals like Shannon entropy and relative entropy. We discuss potential applications of the proposed formalization for the analysis of data compression and security protocols.

1 Introduction

Information theory [19] 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, e.g., [18,5,12].

Traditionally, paper-and-pencil based analytical techniques have been used for information theoretic analysis but these methods do not scale very well to most real-world systems. Therefore, computer simulations are predominantly used for information theoretic analysis these days. However, due to its inherent nature, computer simulation can never ascertain 100% accuracy. This fact is extremely undesirable due to the ever increasing usage of information theoretic analysis in the design of safety and mission critical systems. Formal methods tend to overcome such inaccuracy limitations and therefore a higher-order-logic formalization of information theory has recently been proposed [3]. However, the underlying theories of this development have certain constraints and lack important properties of the quantities formalized, which are necessary for any information theoretic analysis. For example, the theories do not support infinite

M. Van Eekelen et al. (Eds.): ITP 2011, LNCS 6898, pp. 233–248, 2011.

values for functions or integrals, which limits the scope of applications and most importantly prevents the proof of important and necessary theorems, like the Radon Nikodym theorem [7].

This paper is primarily focused towards overcoming these shortcomings as we attempt to raise the state-of-the-art in higher-order-logic theorem proving based information theoretic analysis technique from the existing level, where it is applicable only to isolated facets, to a level allowing formal analysis of contemporary engineering and scientific problems. In this regard, we propose to first develop a rigorous higher-order-logic formalization of measure, probability, Lebesgue integration theories over extended real numbers, which are real numbers extended with positive and negative infinity.

Using extended reals to define the measure theory allows us to work with regular finite non-negative measures, infinite measures as well as signed measures. Working with functions or random variables that can take infinite values, allows us to prove important limiting theorems that are not possible to prove when we do not consider infinite values. In fact, in that case, the limit of a sequence is undefined when the sequence is not convergent. However, in the extended reals case, a limit is always defined and can be infinite. Finally, working with infinite Lebesgue integrals allows us to prove various convergence theorems without requiring the sequences to be convergent.

Building on top of this framework, we formalize Shannon's entropy and the relative entropy, which are most widely used information theoretic principles, and verify their classical properties. In the definition of relative entropy, we need to define the Radon Nikodym derivative and prove its properties. The existence of this derivative for absolutely continuous measures is guaranteed by the so called Radon Nikodym theorem. The proof of this theorem was the main motivation to use the extended reals in the formalization.

All of the above mentioned formalization is done using the HOL theorem prover [8] and the paper provides the associated formalization and verification details. This infrastructure paves the path to the formal information theoretic analysis of many engineering systems and we highlight some of these potential applications of our work in this paper as well.

2 Related Work

Based on the work of Hurd [10] on measure theory, Coble [3] formalized the main concepts of Lebesgue integration and probability and used them in the formalization of information theory in HOL. Coble used this framework to verify anonymity properties of the dining cryptographers protocol. This formalization, however, does not include important convergence theorems and properties of the Lebesgue integral and measurable functions, limiting the scope of its applications. We provided a generalization of this work [13], based on Borel spaces, allowing us to verify those properties and theorems. Both formalizations, however, only consider finite-valued measures, functions and integrals. In this paper, we propose to define a new type for extended reals and use it to formalize measure, Lebesgue integration, probability and main concepts of information theory. Using extended reals in the formalization has many advantages. It allows us to define sigma-finite and other infinite measures as well as signed measures. Properties of the Lebesgue integral like the monotonicity can be proven even for non-integrable functions, but most importantly, it allows us to prove convergence theorems that are valid even for non convergent sequences. The latter was the main reason to define extended-real-valued integrals, to be able to prove the important Radon Nikodym theorem. This theorem, and consequently some of the properties of the Radon Nikodym derivative, could not be proven using the formalizations in [3,13]. The Radon Nikodym derivative is needed in the definition of the relative entropy. To the best our knowledge, this is the first higher-order-logic formalization of these information theoretic notions which also includes their properties.

A formalization of the positive extended reals in HOL was proposed by Hurd [11] and has been imported to the Isabelle theorem prover [16]. We propose a formalization that includes all real numbers as well as the positive infinity $+\infty$ and negative infinity $-\infty$. This has, obviously, the advantage of working with negative extended real numbers, for example for signed measures. A formalization of measure theory defined on the positive extended reals has been developed in Isabelle [9], based on the work of Coble [3]. This has been used to prove the Radon Nikodym theorem. The main difference with our work is the use of extended reals, which allows us to define signed measures as well as have the integral defined on the extended reals for arbitrary functions. Most importantly, in our work, we focus on defining the main concepts of information theory as well as prove their properties. We prove the properties of the Radon Nikodym derivative and use it to define and prove the properties of the relative entropy.

A formalization of the Lebesgue integral on the extended reals has been proposed in Mizar [20]. We provide a more general formalization that allowed us to formalize the Radon Nikodym derivative and prove its properties. To the best of our knowledge, neither the Radon Nikodym derivative and its properties nor the relative entropy have been formalized in Mizar.

3 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 reals, 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.

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 arithmetic operations of $\mathbb{R}$ are extended to $\overline{\mathbb{R}}$ with partial functions. For example the addition is extended as follow.

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

This is formalized in higher-order logic as

```
val extreal_add_def = Define‘
   (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`. Similarly, we extend the other arithmetic operators and prove their properties.

The set of extended real numbers is a totally ordered set such that for all $a \in \overline{\mathbb{R}}$, $-\infty \leq a \leq +\infty$. With this order, $\overline{\mathbb{R}}$ is a complete lattice where every subset has a supremum and an infimum. The supremum is formalized in HOL as:

```
val extreal_sup_def = Define
  ‘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. Next, we tackle the following theorem, which we will use in the Radon Nikodym theorem proof in Section 5

Theorem 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.*

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.

$$x_p(0) = @r.\, r \in P \wedge (\sup P - 1) < r \text{ and}$$
$$x_p(n+1) = @r.\, r \in P \wedge \max(x_p(n), \sup P - \tfrac{1}{2^{n+1}}) < r < \sup P$$

where @ represents the Hilbert choice operator.
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

```
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). 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 in 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 provide an extensive formalization of the extended real numbers, which consists of more than 220 theorems written in around 3000 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. The proof script is available in [14] and can used in a variety of other applications as well. In the next sections, we present the formalization of these theories based on the extended real numbers.

4 Formalization of Measure, Integration and Probability

Using measure theory to formalize probability 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. The expectation of a random variable is its integral with respect to the probability measure. The Lebesgue integral is used because it provides a unique definition for discrete and continuous random variables, it handles a broader class of functions than the Reimann integral, and it exhibits a better behavior when it comes to interchanging limits and integrals. Most of the concepts of this section have already been formalized in HOL [13]. However, the formalization of this paper is based on the extended reals. In this context, the `limit` of a monotonically increasing sequence becomes the `supremum` and can be infinite. This allows us to verify various limiting properties and convergence theorems.

4.1 Measure Theory

By definition, measurable functions satisfy the condition that the inverse image of a measurable set is also measurable, which we formalize in higher-order logic as follows

```
 ⊢ ∀a b f. 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
```

This definition applies to functions defined on arbitrary spaces. We are interested in real-valued measurable functions and hence the Borel sigma algebra on the set of extended real numbers is used. Working with the Borel sigma algebra makes the set of measurable functions a vector space. It also allows us to formally verify various properties of the measurable functions necessary for the formalization of the Lebesgue integral and its properties in HOL.

We define the Borel sigma algebra on $\overline{\mathbb{R}}$, which we call *Borel*, as the smallest sigma algebra generated by the open rays

```
val Borel_def = Define
   ‘Borel = sigma (UNIV:extreal->bool)
                  (IMAGE (λa. {x:extreal | x < a}) UNIV)’
```

where `sigma` is defined as

```
sigma sp st = (sp, ⋂ s | st ⊆ s ∧ sigma_algebra (sp,s))
```

We also prove that the Borel sigma algebra on the extended reals is the smallest sigma algebra generated by any of the following classes of intervals: $[c, +\infty]$, $(c, +\infty]$, $[-\infty, c]$, (c, d), $[c, d)$, $(c, d]$, $[c, d]$, where $c, d \in \mathbb{R}$. Using the above result, we prove that to check the measurability of extended-real-valued function, it is sufficient to check that the inverse image of the open ray is measurable. The same result is valid for the other classes of intervals.

Theorem 2. *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 $\forall c \in \overline{\mathbb{R}},\ f^{-1}([-\infty, c[) \in \mathcal{A}$*

We prove in HOL various properties of the extended-real-valued measurable functions.

- Every constant real function on a space X is measurable.
- The indicator function on a set A is measurable iff A is measurable.
- Let f and g be measurable functions and $c \in \mathbb{R}$, then the following functions are also measurable: $cf, |f|, f^n, f + g, fg$ and $max(f, g)$.
- If (f_n) is a monotonically increasing sequence of real-valued measurable functions such that $\forall\, x,\ f(x) = \sup_{n\in\mathbb{N}} f_n(x)$, then f is a measurable function.

4.2 Lebesgue Integral

The Lebesgue integral is defined using 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, x) as a finite linear combination of indicator functions of measurable sets (a_i) that form a partition of the space X.

$$\forall t \in X,\ g(t) = \sum_{i \in s} x_i I_{a_i}(t) \quad c_i \geq 0 \tag{1}$$

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 positive simple functions then extended to non-negative functions and finally to arbitrary 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 given by

$$\int_X g\, d\mu = \sum_{i \in s} x_i \mu(a_i) \tag{2}$$

This is formalized in HOL as

```
val pos_simple_fn_integral_def = Define
    ‘pos_simple_fn_integral m s a x =
            SIGMA (λi. x i * measure m (a i)) s’
```

While the choice of $((x_i),(a_i),s)$ to represent g is not unique, we prove that the integral as defined above is independent of that choice. We also prove important properties of the Lebesgue integral of positive simple functions such as the linearity and monotonicity. The Lebesgue integral of non-negative measurable functions is given by

$$\int_X f\,d\mu = \sup\{\int_X g\,d\mu \mid g \leq f \text{ and } g \text{ positive simple function}\} \tag{3}$$

Its formalization in HOL is the following

```
val pos_fn_integral_def = Define
    ‘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 arbitrary measurable functions is given by

$$\int_X f\,d\mu = \int_X f^+\,d\mu - \int_X f^-\,d\mu \tag{4}$$

where f^+ and f^- are the non-negative measurable functions defined by $f^+(x) = \max(f(x),0)$ and $f^-(x) = \max(-f(x),0)$.

```
val fn_integral_def = Define
    ‘fn_integral m f = pos_fn_integral m (fn_plus f) -
                       pos_fn_integral m (fn_minus f)’
```

As defined above, the Lebesgue integral can be undefined when the integrals of both f^+ and f^- are infinite. This requires that in most properties of the Lebesgue integral, we assume that the functions are integrable, as defined next.

Definition 1. *Let $(X,\mathcal{A},\mu)$ be a measure space, a measurable function f is integrable iff $\int_X f^+\,d\mu < \infty$ and $\int_X f^-\,d\mu < \infty$*

Lebesgue Monotone Convergence. The 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 [1] and the properties of the integral. We present in the sequel a proof of the theorem in HOL.

Theorem 3. *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*

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

```
 ⊢ ∀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

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

Lebesgue Integral Properties. Most properties of the Lebesgue integral cannot be proved directly from the definition of the integral. We prove instead that any measurable function is the limit of a sequence of positive simple functions. The properties of the Lebesgue integral are then derived from the properties on the positive simple functions.

Theorem 4. *For any non-negative measurable function f there exists a monotonically increasing sequence of positive simple functions (f_n) such that $\forall\, x,\ f(x) = \sup_{n \in \mathbb{N}} f_n(x)$. Besides*

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

The above theorem is formalized in HOL as

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

We prove this theorem by showing that the sequence (f_n), defined below, satisfies the conditions of the theorem and use the Lebesgue monotone convergence theorem to conclude that $\int_X f \, d\mu = \sup_{n \in \mathbb{N}} \int_X f_n \, d\mu$.

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

For arbitrary integrable functions, Theorem 4 is applied to f^+ and f^- and results in a well-defined integral, given by

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

Using Theorem 4, we extend the properties of the Lebesgue integral for positive simple functions to arbitrary integrable functions. The main properties we proved are the monotonicity and linearity of the Lebesgue integral.

4.3 Probability Theory

We formalize the Kolmogorov axiomatic definition of probability using measure theory by defining the sample space Ω, the set F of events which are subsets of Ω and the probability measure p. A probability measure is a measure function and an event is a measurable set. (Ω, F, p) is a probability space iff it is a measure space and $p(\Omega) = 1$. A random variable is by definition a measurable function.

```
val random_variable_def = Define
   ‘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 5. *If X and Y are random variables and $c \in \mathbb{R}$ then the following functions are also random variables: $cX, |X|, X^n, X+Y, XY$ and $max(X,Y)$.*

The probability mass function (PMF) of a random variable X is defined as the function p_X assigning to a set A the probability of the event $\{X \in A\}$. We also formalize the joint probability mass function of two random variables and of a sequence of random variables.

```
val pmf_def = Define
        ‘pmf p X = (λA. prob p (PREIMAGE X A ∩ p_space p))’
```

Finally we use the formalization of the Lebesgue integral to define the expectation of a random variable and its variance. The expectation of a random value X is defined as the integral of X with respect to the probability measure, $E[X] = \int_{\Omega} X \, dp$.

```
val expectation_def = Define ‘expectation = fn_integral’
```

The properties of the expectation are derived from the properties of the integral. The variance of a random variable is defined as $E[|X - E[X]|^2]$. We also prove the properties of the variance in HOL.

5 Measures of Entropy in HOL

In this section, we make use of the formalization of measure, Lebesgue integral and probability theory to formalize fundamental quantities of information theory, namely the Shannon entropy and the relative entropy. We prove some of their properties and present some of their applications. In the definition of relative entropy, we need to define the Radon Nikodym derivative [7] and prove its properties. The existence of this derivative for absolutely continuous measures is guaranteed by the so called Radon Nikodym theorem [7]. The proof of this theorem was the main motivation to use the extended reals in the formalization.

5.1 Shannon Entropy

The Shannon entropy [4] is a measure of the uncertainty associated with a random variable. It is restricted to discrete random variables and its extension to continuous random variables, known as the differential entropy, does not have some of the desired properties. In fact, the differential entropy can be negative and is not invariant under change of variables.

Definition 2. *(Shannon Entropy) The entropy H of a discrete random variable X with alphabet $\mathcal{X}$ and probability mass function p is defined by*

$$H(X) = -\sum_{x \in \mathcal{X}} p(x) log(p(x))$$

We provide, by contrast, a formalization that is based on the expectation and is valid for both discrete and continuous cases. We prove, later, the equivalence between the two definitions.

$$H(X) = E[-log(p(X))]$$

We propose the following formalization of the entropy in higher-order logic.

```
⊢ entropy b p X =  expectation q (λx. - logr b (pmf p X {x}))
```

where, p is the probability space and b is the basis of the logarithm and q is the probability space with respect to which the expectation is defined and is given by

```
⊢  q = (IMAGE X (p_space p), POW (IMAGE X (p_space p)), pmf p X)
```

We then prove the equivalence between the two definitions of entropy, i.e. the expectation based definition and the sum based definition, for the case of a discrete random variable.

```
⊢  entropy b p X = -SIGMA (λx. pmf p X x * logr b (pmf p X {x}))
                                   (IMAGE X (p_space p))
```

We prove the Asymptotic Equipartition Property (AEP) [4] which is the information theoretic analog of the Weak Law of Large Numbers (WLLN) [15]. It states that for a stochastic source X, if its time series $X_1, X_2, \ldots$ is a sequence of independent identically distributed (*iid*) random variables with entropy $H(X)$, then $-\frac{1}{n} log(p(X_1, \ldots, X_n))$ converges in probability to $H(X)$. We prove the AEP by first proving the Chebyshev's inequality and use it to prove the WLLN.

Theorem 6. *(AEP): if $X_1, X_2, \ldots$ are iid then*

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

A consequence of the AEP is the fact that the set of observed sequences, $(x_1, \ldots, x_n)$, for which the 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 holds for the typical sequences is true with high probability and thus determines the average behavior of a large sample.

Definition 3. *(Typical Set) The typical set A_ϵ^n with respect to $p(x)$ is the set of sequences $(x_1, \ldots, x_n)$ satisfying*

$$2^{-n(H(X)+\epsilon)} \leq p(x_1, \ldots, x_n) \leq 2^{-n(H(X)-\epsilon)}$$

We use the AEP to prove that the typical set has a total probability equal to 1 and that the total number of typical sequences is upper bounded by $2^{n(H(X)+\epsilon)}$.

5.2 Relative Entropy

The relative entropy [4] or Kullback Leibler divergence $D(\mu||\nu)$ is a measure of the distance between two distributions μ and ν. It is defined as

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

where $\frac{d\nu}{d\mu}$ is the Radon Nikodym derivative of ν with respect to μ. This derivative is a non-negative measurable function that, when it exists, satisfies for any measurable set.

$$\int_A \frac{d\nu}{d\mu}\, d\mu = \nu(A)$$

The Radon Nikodym derivative is formalized in HOL as

```
val RN_deriv_def = Define
   ‘RN_deriv m v =
     @f. f IN measurable (m_space m, measurable_sets m) Borel ∧
     (∀a. a ∈ measurable_sets m ⇒
      (fn_integral m (λx. f x * indicator_fn a x) = measure v a))’
```

The relative entropy is then formalized as

```
val KL_divergence_def = Define
   ‘KL_divergence b m v =
      - fn_integral m (λx. logr b ((RN_deriv m v) x))’
```

The existence of the Radon Nikodym derivative is guaranteed for absolutely continuous measures by the Radon Nikodym theorem. A measure ν is absolutely continuous with respect to the measure μ iff for every measurable set A, $\mu(A) = 0$ implies that $\nu(A) = 0$. Next, we state and prove the Radon Nikodym theorem (RNT) for finite measures. The theorem can be easily generalized to sigma finite measures.

Theorem 7. *(RNT) If ν is absolutely continuous with respect to μ, then there exists a non-negative $\mu-$integrable function f such that for any measurable sets,*

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

The Radon Nikodym theorem is formalized in HOL as follows,

```
⊢ ∀m v. measure_space m ∧ measure_space v ∧
 (m_space v = m_space m) ∧
 (measurable_sets v = measurable_sets m) ∧
 (measure_absolutely_continuous m v) ∧
 (measure v (m_space v) ≠ PosInf) ∧
 (measure m (m_space m) ≠ PosInf) ⇒
 (∃f. f ∈ measurable (m_space m,measurable_sets m) Borel ∧
 (∀A. A ∈ measurable_sets m ⇒
  (pos_fn_integral m (λx. f x * indicator_fn A x) = measure v A)))
```

To prove the theorem, we prove the following lemma, which we propose as a generalization of Theorem 1. To the best of our knowledge, this lemma has not been referred to in textbooks and we find that it is a useful result that can be used in other proofs.

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*

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

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 prove that the function g, defined below, satisfies the conditions of the theorem.

$$\forall x. \; g(x) = \sup_{n \in \mathbb{N}} f_n(x)$$

The main reason we used the extended reals in our formalization was the inability to prove the Radon Nikodym theorem without considering infinite values. In fact, in our proof, we use the Lebesgue monotone convergence to prove that

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

However, the Lebesgue monotone convergence in [13] which does not support the extended reals, requires the sequence f_n to be convergent, which is not necessarily the case here and cannot be added as an assumption because the sequence f_n is generated inside the proof. The Lebesgue monotone convergence theorem with the extended reals is valid even for sequences that are not convergent since it uses the `sup` operator instead of the limit `lim`.

Next, we prove the following properties of the Radon Nikodym derivative.

- The Radon Nikodym derivative of ν with respect to μ is unique, μ almost-everywhere, i.e., unique up to a null set with respect to μ.
- If ν_1 and ν_2 are absolutely continuous with respect to μ, then $\frac{d(\nu_1+\nu_2)}{d\mu} = \frac{d\nu_1}{d\mu} + \frac{d\nu_2}{d\mu}$, μ almost-everywhere.
- If ν is absolutely continuous with respect to μ and $c \geq 0$, then $\frac{d(c*\nu)}{d\mu} = c*\frac{d\nu}{d\mu}$, μ almost-everywhere.

For finite spaces, we prove the following two results for the Radon Nikodym derivative and the relative entropy.

$$\forall x \in X, \mu\{x\} \neq 0 \Rightarrow \frac{d\nu}{d\mu}(x) = \frac{\nu\{x\}}{\mu\{x\}}$$

$$\forall x \in X, \nu\{x\} \neq 0 \Rightarrow D(\mu||\nu) = \sum_{x\in X} \mu\{x\} \log \frac{\mu\{x\}}{\nu\{x\}}$$

Finally, the relative entropy between the joint distribution $p(x, y)$ of two random variables X and Y and the product of their marginal distributions $p(x)$ and $p(y)$ is equal to the mutual information $I(X, Y)$.

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

5.3 Applications

The developed formalization of entropy measures can be used in a number of engineering applications. For instance, the formally verified AEP and the typical set, formalized in Section 5.1, can be directly applied in the proof of the Shannon source coding theorem which establishes the fundamental limit of data compression. It 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 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.

One way to prove the above theorem is to propose an encoding scheme that is based on the typical set. The average codeword length for all sequences is close to the average codeword length considering only the typical sequences, because, asymptotically, the total probability of the typical set is equal to 1. From the upper bound on the number of typical sequences, we deduce that the average number of bits needed to encode the typical sequences can be made arbitrarily close to $nH(X)$.

Quantitative theories of information flow are gaining a lot of attention in a variety of contexts, such as secure information flow, anonymity protocols, and side-channel analysis. Various measures are being proposed to quantify the flow of information. Serjantov [18] and Diaz et al. [6] 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.

Deng [5] proposed the relative entropy as a measure of the amount of information revealed to the attacker after observing the outcomes of the protocol, together with the apriori information. We can use our formalization of the relative entropy developed in Section 5.2 to apply this technique to verify the anonymity properties of the Dining Cryptographers [2] and Crowds [17] protocols.

6 Conclusions

In this paper, we have presented a formalization in HOL of measure, Lebesgue integration and probability theories defined on the extended reals. We used this infrastructure, to formalize main concepts of information theory, namely the Shannon entropy and relative entropy. The formalization based on the extended reals enables us to verify important properties and convergence theorems as well as prove the important Radon Nikodym theorem. The latter allows us to prove the properties of the Radon Nikodym derivative, used in the definition of the relative entropy.

The verification of properties of the Shannon entropy and relative entropy makes it possible to perform information theoretic analysis on a wide range of applications. Using our formalization, we proved the Asymptotic Equipartition Property in HOL, which is used to define and verify the notion of typical sets. This, in turn, is the basis to prove the Shannon source coding theorem, providing the fundamental limits of data compression. The relative entropy is an important measure of divergence between probability distributions. It is used to define other concepts of information theory, but it is also used in several other applications like the anonymity application in [5].

Our future work include applying the technique in [5] to verify the anonymity properties of the Dining Cryptographers and Crowds protocols within the sound core of a theorem prover. We also plan to work out the details of the applications outlined in Section 5.3.

The HOL code for the formalization presented in this paper is available in [14]. It required more than 11000 lines of code and contains around 500 theorems. Most of this formalization is very generic and thus can be utilized to formalize more advanced mathematics or formally reason about a more wide range of engineering applications.

References

1. Bogachev, V.I.: Measure Theory. Springer, Heidelberg (2006)
2. Chaum, D.: The Dining Cryptographers Problem: Unconditional Sender and Recipient Untraceability. Journal of Cryptology 1(1), 65–75 (1988)
3. Coble, A.R.: Anonymity, Information, and Machine-Assisted Proof. PhD thesis, University of Cambridge (2010)
4. Cover, T.M., Thomas, J.A.: Elements of Information Theory. Wiley Interscience, Hoboken (1991)
5. Deng, Y., Pang, J., Wu, P.: Measuring Anonymity with Relative Entropy. In: Dimitrakos, T., Martinelli, F., Ryan, P.Y.A., Schneider, S. (eds.) FAST 2006. LNCS, vol. 4691, pp. 65–79. Springer, Heidelberg (2007)
6. Díaz, C., Seys, S., Claessens, J., Preneel, B.: Towards Measuring Anonymity. In: Dingledine, R., Syverson, P.F. (eds.) PET 2002. LNCS, vol. 2482, pp. 54–68. Springer, Heidelberg (2003)
7. Goldberg, R.R.: Methods of Real Analysis. Wiley, Chichester (1976)
8. Gordon, M.J.C., Melham, T.F.: Introduction to HOL: A Theorem Proving Environment for Higher-Order Logic. Cambridge University Press, Cambridge (1993)
9. Hölzl, J.: Mechanized Measure, Probability, and Information Theory. Technical University Munich, Germany (2010), http://puma.in.tum.de/p-wiki/images/d/d6/szentendre_hoelzl_probability.pdf
10. Hurd, J.: Formal Verifcation of Probabilistic Algorithms. PhD thesis, University of Cambridge (2002)
11. Hurd, J., McIver, A., Morgan, C.: Probabilistic Guarded Commands Mechanized in HOL. Electronic Notes in Theoretical Computer Science 112, 95–111 (2005)
12. Malacaria, P.: Assessing Security Threats of Looping Constructs. SIGPLAN Not. 42(1), 225–235 (2007)
13. Mhamdi, T., Hasan, O., Tahar, S.: On the formalization of the lebesgue integration theory in HOL. In: Kaufmann, M., Paulson, L.C. (eds.) ITP 2010. LNCS, vol. 6172, pp. 387–402. Springer, Heidelberg (2010)
14. Mhamdi, T., Hasan, O., Tahar, S.: Formalization of Measure and Lebesgue Integration over Extended Reals in HOL. Technical Report, ECE Dept., Concordia University (February 2011), http://hvg.ece.concordia.ca/Publications/TECH_REP/MLX_TR11/
15. Papoulis, A.: Probability, Random Variables, and Stochastic Processes. Mc-Graw Hill, New York (1984)
16. C. Paulson, L.: Isabelle: a Generic Theorem Prover. Springer, Heidelberg (1994)
17. Reiter, M.K., Rubin, A.D.: Crowds: Anonymity for Web Transactions. ACM Transactions on Information and System Security 1(1), 66–92 (1998)
18. Serjantov, A., Danezis, G.: Towards an Information Theoretic Metric for Anonymity. In: Dingledine, R., Syverson, P.F. (eds.) PET 2002. LNCS, vol. 2482, pp. 41–53. Springer, Heidelberg (2003)
19. Shannon, C.E.: A Mathematical Theory of Communication. The Bell System Technical Journal 27(3), 379–423 (1948)
20. Shidama, Y., Endou, N., Kawamoto, P.N.: On the formalization of lebesgue integrals. Studies in Logic, Grammar and Rhetoric 10(23), 167–177 (2007)