Formalization of DPI and Jensen's Inequality in HOL

Cvetan Dunchev, Ghassen Helali, Osman Hasan and Sofiène Tahar

Department of Electrical and Computer Engineering,
Concordia University, Montreal, QC, Canada
{dunchev,helali,o_hasan,tahar}@ece.concordia.ca

Technical Report

April 2016

Abstract

Nowadays, protecting the sensitive information is becoming more important. Many applications are subject to the flow of secret data that need to be kept hidden. In this context, quantitative analysis of information flow is a promising technique to reason about security aspects of these systems. Various information theory properties can be verified using probabilistic analysis. In this report, we present a formal verification of the Data Processing Inequality (DPI) and the Jensen's Inequality in the higher-order logic theorem prover HOL4. While the DPI property states that no clever manipulation of a given data can improve the inferences that can be made from the data, Jensen's inequality in its general definition is a relation between the value of a convex function of the integral and the integral of a convex function. Due to its generality, Jensen's inequality has many applications in the information theory.

1 Introduction

Quantitative analysis of information flow is a very important technique in order to determine the security and the robustness of a communication system. The quantitative analysis is mostly conducted using probability theory and its related properties in order to reason about the information flow properties.

In Information theory, the mutual information [4] captures the quantity of information that a random variable has about another random variable. It is measured usually in units of bits and can be considered as the reduction in uncertainty about a random variable given the knowledge of another random variable. Let X and Y be random variables, the *mutual information* $I(X;Y)$ for the discrete case is defined to be:

$$\sum_{x,y} P_{XY}(x,y).log\frac{P_{XY}(x,y)}{P_X(x).P_Y(y)}$$

where $P_{XY}(x, y)$ is the joint distribution of X and Y, $P_X(x)$ and $P_Y(y)$ are the marginal distributions of X and Y, respectively.

In order to better understand the notion of mutual information, we should first introduce the Shannon *entropy*. Informally, entropy represents the average amount of an information an information flow within a communication system; but formally, it is the measure of uncertainty of a random variable. The higher the entropy is, the more uncertain we are about the random variable. The entropy $H(X)$ of a random variable X for the discrete probability space is defined as:

$$-\sum_{x} P_X(x).logP_X(x)$$

The entropy can be naturally extended to the notion of *conditional entropy*, $H(X|Y)$, which is the average uncertainty about a random variable X after observing another random variable Y. From the notions described above, we can derive numerous information flow properties. The Data Processing Inequality (DPI) is a very useful property in many applications in the case of any three random variables X, Y and Z such that the conditional distribution of Z depends only on Y and is conditionally independent of X. We say that X, Y and Z in this order satisfy the *Markov property* if and only if X and Z are conditionally independent given Y, i.e., $p(x, z|y) = p(x|y).p(z|y)$, where $p(x)$ is the usual abbreviation of $P_X(X = x)$. In this case we say that X, Y and Z form a *Markov chain*, $X \rightarrow Y \rightarrow Z$. The Markov property will play a key role in the formalization of the DPI as it will be described in the next sections. Jensen's inequality relates the value of a convex function of the integral to the integral of a convex function. Jensen's inequality is a very general property that can be applied in many fields of information theory. In this report, we will investigate the formalization of both notions. In this context, we propose to use the formalized probability theory in higher-order logic theorem prover HOL4 [1] in order to formally verify the DPI and Jensen's inequality.

2 Probability and Information Theories

Probability and information theories provide mathematical models to evaluate the uncertainty of random phenomena. These concepts are commonly used in different fields of engineering and computer science, such as signal processing, data compression and data communication, to quantify the information. Recently, the probability and information theories have been widely used for cryptographic and information flow analysis [12]. Some foundational notions of these formalizations are described below.
Let X and Y denote discrete random variables, with x and y, and $\mathcal{X}$ and $\mathcal{Y}$ denoting their specific values and set of all possible values, respectively. Similarly, the probability of X and Y being equal to x and y is denoted by $p(x)$ and $p(y)$, respectively.

- Probability Space: *a measure space such that the measure of the state space is* 1
- Independent Events: *Two events X and Y are independent iff $p(X \cap Y) = p(X)p(Y)$.*
- Random Variable: *$X : \Omega \rightarrow \mathcal{R}$ is a random variable iff X is $(F, \mathcal{B}(\mathcal{R}))$ measurable where F denotes the set of events and $\mathcal{B}$ is the Borel sigma algebra.*
- Joint Probability: *A probabilistic measure where the likelihood of two events occurring together and at the same point in time is calculated. Joint probability is the probability of event Y occurring at the same time event X occurs. It is mathematically expressed as $p(X \cap Y)$ or $p(X, Y)$.*

- Conditional Probability: *A probabilistic measure where an event X will occur, given that one or more other events Y have occurred. Mathematically* $p(X|Y)$ *or* $\frac{p(X \cap Y)}{p(X)}$.
- Expected Value: $E[X]$ *of a random variable X is its Lebesgue integral with respect to the probability measure. The following properties of the expected value have been verified in HOL4* [10]:
  1. $E[X + Y] = E[X] + E[Y]$
  2. $E[aX] = aE[X]$
  3. $E[a] = a$
  4. $X \leq Y$ then $E[X] \leq E[Y]$
  5. X and Y are independent then $E[XY] = E[X]E[Y]$
- Variance and Covariance: *Variance and covariance have been formalized in HOL4 using the formalization of expectation. The following properties have been verified* [10]:
  1. $Var(X) = E[X^2] - E[X]^2$
  2. $Cov(X, Y) = E[XY] - E[X]E[Y]$
  3. $Var(X) \geq 0$
  4. $\forall a \in R, Var(aX) = a^2 Var(X)$
  5. $Var(X + Y) = Var(X) + Var(Y) + 2Cov(X, Y)$

The above-mentioned definitions and properties have been utilized to formalize the foundations of information theory in HOL4 [10]. The widely used information theoretic measures can be defined as:

- The Shannon Entropy: *It measures the uncertainty of a random variable*

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

- The Conditional Entropy: *It measures the amount of uncertainty of X when Y is known*

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

- The Mutual Information: *It represents the amount of information that has been leaked*

$$I(X;Y) = I(Y;X) = H(X) - H(X|Y)$$

- The Relative Entropy or Kullback Leiber Distance: *It measures the inaccuracy or information divergence of assuming that the distribution is q when the true distribution is p*

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

- The Guessing Entropy: *It measures the expected number of tries required to guess the value of X optimally*

$$G(X) = \sum_{1 \leq i \leq n} ip(x_i)$$

- The Rényi Entropy: *It is related to the difficulty of guessing the value of X*

$$H_\alpha(X) = \frac{1}{1-\alpha} log\,(\sum_{x \in \mathcal{X}} P[X = x]^\alpha)$$

Among the measures listed above, Mhamdi [8] and Coble [3] formalized the Entropy, Conditional Entropy, Relative Entropy and Mutual Information in HOL4 and Hölzl [6] formalized similar concepts in Isabelle/HOL.

3 Shannon Based Information Flow

In this section, we review the formalization of the most common measures to quantify information flow, such as Shannon entropy, relative entropy and mutual information [11]. These measures will be then used to formally verify the Data Processing Inequality as well as Jensen's Inequality properties.

3.1 Shannon Entropy

Shannon entropy measures are considered the most common measures of information theory. These measures have been formalized in [10] using the formalized foundations of measure, Lebesgue integral and probability theories [9]. In order to formalize the relative entropy, the Radon Nikodym derivative has been defined and the related properties have been verified in HOL4. The Shannon entropy measures the uncertainty of a random variable and is defined as follows

Definition 1 (Shannon Entropy).
The entropy H of a random variable X with alphabet $\mathcal{X}$ and probability mass function p is defined by

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

which is equivalent to the following expectation-based definition

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

The corresponding formalization in higher-order logic is

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

Many related properties have been formally verified, such as the Asymptotic Equipartition Property (AEP) which is the equivalent of the Weak Law of Large Numbers (WLLN) [16].

3.2 Relative Entropy

The Kullback Leibler (KL) divergence, also called relative entropy, measures the distance between two distributions p and q.

Definition 2 (Relative Entropy).
The relative entropy of two distributions p and q is

$$\mathcal{D}(p||q) \;=\; -\int_X log\ \frac{dp}{dq} dq$$

where $\frac{dp}{dq}$ is the Radon Nikodym derivative of p with respect to q. The Kullback Leibler divergence is formalized in HIOL4 as

```
⊢ KL_divergence b p q = - fn_integral p (λx.  logr b ((RN_deriv p q) x))
```

3.3 Mutual Information

The mutual information measures the dependance between the two random variables. It describes the uncertainty about one of them such that the other is known. Mutual information is considered as a measure of information leakage. In [8], it is formalized as the KL divergence between the joint distribution and the product of marginal distributions.

Definition 3 (Mutual Information).
$I(X,Y) \;=\; D(p(x,y)||p(x)p(y)) \;=\; sum_{(x,y)} p(x,y) log \frac{p(x,y)}{p(x)p(y)}$

This definition is then formalized in HOL4 as follows

```
⊢mutual_information b p s1 s2 X Y =
   let prod_space =
     prod_measure_space (space s1, subsets s1, distribution p X)
       (space s2, subsets s2, distribution p Y) in
     KL_divergence b (p_space prod_space, events prod_space)
       (joint_distribution p X Y) (prob prod_space)
```

The detailed formalization of the quantities described above and their related properties is provided in [8].

4 Data Processing Inequality

Our formalization builds on top of the work done by Liu [7] and Mhamdi [14] who formalized the DTMC and probability theory, respectively. We start our formalization by defining the Markov property in HOL4. It is a direct consequence of the definition of the Markov property in [7]:

Definition 4 (Markov Chain).

```
⊢ mc p X Y Z = (∀x y z.  joint_distribution3 p X Y Z (x,y,z) =
  distribution p X x * (conditional_distribution p Y X y x) *
   (conditional_distribution p Z Y z y))
```

The motivation behind this definition relies on the fact that if the random variables X, Y and Z in this order satisfy the Markov property, then the joint distribution $p(x,y,z)$ is equal to $p(x).p(y|x).p(z|y)$. Indeed, using the conditional distribution we have:

$$p(x,y,z) = p(x).p(y,z|x) = p(x).p(y|x).p(z|x,y)$$

and using the Markov property [7], $p(z|x,y) = p(z|y)$ holds. The difinitions of `distribution`, `joint_distribution3` and `conditional_distribution` are straightforward and can be found in [2]. For instance, `joint_distribution3` refers to the joint distribution of 3 random variables, i.e., $P(X,Y,Z)$. These definitions can be also useful for any other probabilistic analysis. They are also a part of the latest HOL4 distribution [2].

Next, we give the definition of the mutual information $I(X;Y)$ between the random variables X and Y:

Definition 5 (Mutual Information).

```
⊢ mutual_information b p s1 s2 X Y = (let prod_space =
        prod_measure_space (space s1, subsets s1, distribution p X)
          (space s2, subsets s2, distribution p Y) in
          KL_divergence b (p_space prod_space, events prod_space)
          (joint_distribution p X Y) (prob prod_space))
```

The formalization of the Kullback-Leibler Divergence, `KL_divergence`, $D_{KL}(P_X||P_Y)$, of the difference between the probability distributions X and Y over the probability spaces s_1 and s_2, respectively, can be found in [13]. In order to prove the DPI, we first have to prove a number of intermediate lemmas. A first result is to rewrite the mutual information as a sum of distributions which is formalized in HOL as follows:

Theorem 1 (Mutual Information as a Sum of Distributions).

```
⊢ ∀b p X Y Z.(POW (p_space p) = events p) ∧
   random_variable X p (IMAGE X (p_space p), POW (IMAGE X (p_space p))) ∧
   random_variable Y p (IMAGE Y (p_space p), POW (IMAGE Y (p_space p))) ∧
   random_variable Z p (IMAGE Z (p_space p), POW (IMAGE Z (p_space p))) ∧
   FINITE (p_space p) ⇒

 mutual_information b p (IMAGE X (p_space p), POW (IMAGE X (p_space p)))
    (IMAGE Y (p_space p) × IMAGE Z (p_space p),
     POW (IMAGE Y (p_space p) × IMAGE Z (p_space p))) X (λ x.(Y x,Z x)) =
 Normal (SIGMA (λ(x,(y,z)).(joint_distribution p X (λx.(Y x,Z x)) (x,(y,z)))
     * logr b ((joint_distribution p X (λ x.(Y x,Z x)) (x,(y,z)))/
     ((distribution p X x) * (distribution p (λ x.(Y x,Z x)) (y,z)))))
     IMAGE X (p_space p) × (IMAGE Y (p_space p) × IMAGE Z (p_space p)))
```

where `POW` is the power set function and `IMAGE f s` refers to the image of a set s with the function f.

The above statement is the formalization of the property reducing the mutual information to a sum of distributions, mathematically this property is defined as follow:

$$I(X,\ (Y,\ Z)) \ =\ \sum_{(x,(y,z))} P(X,\ (Y,\ Z))\ .\ \frac{log_b\ (P(X,\ (Y,\ Z)))}{P(x)\ .\ P(Y,\ Z)}$$

This theorem can be found under `finite_mutual_information_reduce2` in [15]. Since the `KL_divergence` is defined over the extended reals [13], it is of type `extreal` and therefore the mutual information has the same type. We therefore use the function `Normal` [13] which converts an object from type `real` to object of type `extreal`. Having the mutual information, the next step in our formalization is the conditional mutual information $I(X;Y|Z)$ of X and

Y given Z, which is defined as $I(X;Y\times Z)-I(X;Z)$, where $Y\times Z$ is the random variable $\lambda x.(Y(x),Z(x))$:

Definition 6 (Conditional Mutual Information).

```
⊢ conditional_mutual_information b p s1 s2 s3 X Y Z =
    let prod_space =
       prod_measure_space (space s2, subsets s2, distribution p Y)
         (space s3, subsets s3, distribution p Z) in
       mutual_information b p s1 (p_space prod_space, events prod_space) X
         (λ x.(Y x, Z x)) - mutual_information b p s1 s3 X Z
```

Using the commutativity of the distribution function which says that $P_{Y\times Z}((y,z))=P_{Z\times Y}((z,y))$, we have the following equality: $I(X;Y\times Z)=I(X;Z\times Y)$. Therefore, we have the equality:

$$I(X;Y|Z)+I(X;Z)=I(X;Y\times Z)=I(X;Z\times Y)=I(X;Z|Y)+I(X;Y)$$

which is formalized as follows:

Theorem 2 ($I(X;Y|Z)+I(X;Z)=I(X;Z|X)+I(X;Y)$).

```
⊢ ∀b p X Y Z.(POW (p_space p) = events p) ∧ prob_space p ∧
  (s1 = (IMAGE X (p_space p),
         POW (IMAGE X (p_space p)))) ∧
  (s2 = (IMAGE Y (p_space p), POW (IMAGE Y (p_space p)))) ∧
  (s3 = (IMAGE Z (p_space p), POW (IMAGE Z (p_space p)))) ∧
  (s23 = (IMAGE (λ x.(Y x, Z x)) (p_space p),
         POW (IMAGE (λ x.(Y x, Z x)) (p_space p)))) ∧
  (s32 = (IMAGE (λ x.(Z x, Y x)) (p_space p),
         POW (IMAGE (λ x.(Z x, Y x)) (p_space p)))) ∧
  random_variable X p s1 ∧ random_variable Y p s2 ∧ random_variable Z p s3 ∧
  random_variable (λ x.(Y x, Z x)) p s23 ∧
  random_variable (λ x.(Z x, Y x)) p s32 ∧ FINITE (p_space p) ∧
  mutual_information b p s1 s2 X Y ≠ -∞ ∧
  mutual_information b p s1 s2 X Y ≠ +∞ ∧
  mutual_information b p s1 s3 X Z ≠ -∞ ∧
  mutual_information b p s1 s3 X Z ≠ +∞ ⇒

 conditional_mutual_information b p s1 s3 s2 X Z Y +
   mutual_information b p s1 s2 X Y =
 conditional_mutual_information b p s1 s2 s3 X Y Z +
   mutual_information b p s1 s3 X Z
```

The proof of the above theorem relies on the proof the following property(Theorem 3). It holds because of the associativity of the join distribution, namely $P_{X\times Y\times Z}(x,(y,z))=P_{X\times Z\times Y}(x,(z,y))$ and the symmetry of the additivity:

Theorem 3 (Symmetry of the Additivity).

```
⊢ ∀ x y z.((joint_distribution p X (λ x.(Z x,Y x)) (x,z,y) =
     joint_distribution p X (λ x.(Y x,Z x)) (x,y,z)) ∧
   distribution p (λ x.(Y x,Z x)) (y,z) = distribution p (λ x.(Z x,Y x)) (z,y))
```

```
⇒
 SIGMA (λ (x,z,y).joint_distribution p X (λ x.(Z x,Y x)) (x,z,y) *
   logr b (joint_distribution p X (λ x.(Z x,Y x)) (x,z,y) /
    (distribution p X x * distribution p (λ x.(Z x,Y x)) (z,y))))
   (IMAGE X (m_space p) × (IMAGE Z (m_space p) × IMAGE Y (m_space p)))
    =
 SIGMA (λ (x,y,z).joint_distribution p X (λ x.(Y x,Z x)) (x,y,z) *
   logr b (joint_distribution p X (λ x.(Y x,Z x)) (x,y,z) /
    (distribution p X x * distribution p (λ x.(Y x,Z x)) (y,z))))
   (IMAGE X (m_space p) × (IMAGE Y (m_space p) × IMAGE Z (m_space p)))
```

Furthermore, we proved the more general case:

Theorem 4 (Symmetry of the Additivity (General)).

```
∀f s1 s2 s3.  FINITE s1 ∧ FINITE s2 ∧ FINITE s3
⇒
(SIGMA (λ(x,y,z).f (x,y,z)) (s1 × (s2 × s3)) =
 SIGMA (λ(x,z,y).f (x,y,z)) (s1 × (s3 × s2)));
```

In order to verify the main goal of our work, namely the DPI which says that if $X \to Y \to Z$, then $I(X;Z) \leq I(X;Y)$, we have to verify the following two properties:
Properties.

- **P1:** For all random variables X and Y, the mutual information between X and Y is non-negative, i.e., $\forall X, Y.\ I(X;Y) \geq 0$
- **P2:**If $X \to Y \to Z$, then $I(X;Z|Y) = 0$

  Applying the above two properties to the equality

$$I(X;Y|Z) + I(X;Z) = I(X;Z|Y) + I(X;Y)$$

proves the DPI: $I(X;Z) \leq I(X;Y)$. The first property (P1) is formalized in HOL4 as follows:

Theorem 5 (Positive Mutual Information).

```
⊢ ∀b p X Y.(1 ≤ b) ∧ (POW (p_space p) = events p) ∧
    random_variable X p (IMAGE X (p_space p), POW (IMAGE X (p_space p))) ∧
    random_variable Y p (IMAGE Y (p_space p), POW (IMAGE Y (p_space p))) ∧
    FINITE (p_space p) ⇒
  0 ≤ (mutual_information b p (IMAGE X (p_space p), POW (IMAGE X (p_space p)))
     (IMAGE Y (p_space p), POW (IMAGE Y (p_space p))) X Y)
```

The second statement (P2) is considered as a key lemma in the proof and it is formalized as follows:

Theorem 6.

```
∀ X Y Z. prob_space p ∧ POW (p_space p) = events p ∧
    ∀x.(distribution p X x ≠ 0) ∧
```

```
    ∀z y.  conditional_distribution p Z Y z y ≠ 0 ∧
    ∀b.  joint_distribution3_lambda_def b p X Z Y ∧
    ∀x z y.joint_distribution3 p X Z Y (x,z,y) ≠ 0 ∧
    joint_distribution3_lambda b p X Z Y ∧
    IMAGE X (p_space p) × IMAGE (λ x.(Z x,Y x)) (p_space p) =
      {(a,b,c)|(a ∈ IMAGE X (p_space p) ∧ b ∈ IMAGE Z (p_space p) ∧
        c ∈ IMAGE Y (p_space p))} ∧
    s1 = (IMAGE X (p_space p), POW (IMAGE X (p_space p))) ∧
    s2 = (IMAGE Y (p_space p), POW (IMAGE Y (p_space p))) ∧
    s3 = (IMAGE Z (p_space p), POW (IMAGE Z (p_space p))) ∧
    random_variable X p s1 ∧ random_variable Y p s2 ∧ random_variable Z p s3 ∧
    FINITE (p_space p) ∧ mc p X Y Z ∧ FINITE (IMAGE X (p_space p)) ∧
    FINITE (IMAGE Y (p_space p)) ∧ FINITE (IMAGE Z (p_space p)) ∧
    (p_space (prod_measure_space (space s3,subsets s3,distribution p Z)
      (space s2,subsets s2,distribution p Y)) =
     IMAGE (λ x.(Z x, Y x)) (p_space p)) ∧
    (events (prod_measure_space (space s3,subsets s3,distribution p Z)
      (space s2,subsets s2,distribution p Y)) =
     POW (IMAGE (λ x.(Z x, Y x)) (p_space p))) ∧
    random_variable (λ x.(Z x, Y x)) p (IMAGE (λ x.(Z x, Y x)) (p_space p),
      POW (IMAGE (λ x.(Z x, Y x)) (p_space p))) ∧ (sigma2to3 p X Y Z b) ∧
    (∀x y.joint_cond p X Y x y) ∧ (∀y z.joint_cond p Y Z y z)) ⇒

    conditional_mutual_information b p s1 s3 s2 X Z Y = 0
```

After several rewritings, the lemma is reduced to the following statement:

Theorem 7. *(Theorem 6, contd.)*

```
⊢ SIGMA (λ(x,y).joint_distribution p X (λx.(Z x,Y x)) (x,y) *
    logr b (joint_distribution p X (λx.(Z x,Y x)) (x,y) /
      (distribution p X x * distribution p (λx.(Z x,Y x)) y)))
    {IMAGE X (p_space p) × IMAGE Z (p_space p) × IMAGE Y (p_space p)
=
    SIGMA (λ(x,y).joint_distribution p X Y (x,y) *
      logr b (joint_distribution p X Y (x,y) /
        (distribution p X x * distribution p Y y)))
    IMAGE X (p_space p) × IMAGE Y (p_space p)
```

In Theorems 6 and 7, the Markov property plays its crucial role. Therefore, the left-hand side of the equality formalized in Theorem 7 can be simplified as follows:

$$log_b \frac{p(x,y,z)}{p(x).p(z,y)} = log_b \frac{p(x).p(y|x).p(z|y)}{p(x).p(z,y)} = log_b \frac{p(y|x).p(z|y)}{p(z,y)} = log_b \frac{p(y|x).p(z|y)}{p(y).p(z|y)} = log_b \frac{p(y|x)}{p(y)}$$

The right-hand side of the equality of the same theorem can be rewritten as follows:

$$log_b \frac{p(x,y).p(x)}{p(y)} = log_b \frac{p(x).p(y|x)}{p(x).p(y)} = log_b \frac{p(y|x)}{p(y)}$$

In HOL4 we formalize this lemma as follows:

Theorem 8. *(Intermediate Result for Theorem 6)*

```
⊢ ∀ X Y Z x y z b.  mc X Y Z ⇒
   logr b (joint_distribution p X (λx.(Z x,Y x)) (x,y,z) /
    (distribution p X x * distribution p (λx.(Z x,Y x)) (z,y))))
=
   logr b (joint_distribution p X Y (x,y) /
    (distribution p X x * distribution p Y y)))
```

Using the property above, our main lemma reduces to a more general theorem. Thanks to the Markov property, we substitute both logarithms in the equality with a function $f(x, y)$:

Theorem 9. *(Subgoal of Theorem 6)*

```
⊢ ∀p X Y Z f.  prob_space p ∧ FINITE (p_space p) ∧
   (events p = POW (p_space p)) ∧
   FINITE (IMAGE X (p_space p) × IMAGE Y (p_space p)) ⇒

(SIGMA (λ((x,y),z).joint_distribution p (λx.(X x,Y x)) Z ((x,y),z) * f(x,y))
    IMAGE X (p_space p) × IMAGE Y (p_space p) × IMAGE Z (p_space p) =
  SIGMA (λ(x,y).distribution p (λx.(X x,Y x)) (x,y) * f(x,y) )
    IMAGE X (p_space p) × IMAGE Y (p_space p))
```

5 Jensen's Inequality

Jensen's inequality has applications in many fields of applied mathematics and specifically in information theory. For example, it plays a key role in the proof of the *Information inequality*, $0 \leq D(p||q)$ [4]. We proved Jensen's inequality in its measure theoretic form. We start with the definition of the convex function. If f is a continuous function and $x_1 < x_2 < x_3$, then f is called *convex* iff:

$$\frac{f(x_2)-f(x_1)}{x_2-x_1} = \frac{f(x_3)-f(x_2)}{x_3-x_2}$$

In HOL4 we define the convex function as follows:

Definition 7 (Convex function).

```
⊢ conv_func f =
   ∀ x y z:real.  (x<y ∧ y<z) ⇒ ((f(y)-f(x))/(y-x) ≤ (f(z)-f(y))/(z-y))
```

Now, let Ω be a probability space, μ is a measure funcion on Ω, g and f be an arbitrary and a convex functions on real numbers, respectively. The following statement is known as *Jensen's inequality*:

$$\int_{\Omega} f(g(x))\,\mathrm{d}\mu \geq f(\int_{\Omega} g(x)\,\mathrm{d}\mu)$$

The most challenging part of the proof of Jensen's inequality is to prove the existence of sub-derivatives a and b of f, such that for all x, $a.x + b \leq f(x)$, where for $x_0 = \int_\Omega g(x)\,\mathrm{d}\mu$ we reach the equality $a.x_0 + b = f(x_0)$. This follows from the following two facts:

- according to the Mean value theorem, there exists ν such that if $x < \nu < \xi$, then:

$$\frac{f(x)-f(\xi)}{x-\xi} = f'(\nu)$$

- since f is convex, then its derivative increases, i.e., $f'(\nu) \leq f'(\xi)$

Having a and b, the proof of Jensen's inequality is straightforward:

$$\int_\Omega f(g(x))\,\mathrm{d}\mu \geq \int_\Omega (a.g(x)+b)\,\mathrm{d}\mu = a.\int_\Omega g(x)\,\mathrm{d}\mu + b.\int_\Omega 1\,\mathrm{d}\mu = a.x_0 + b = f(x_0) = f(\int_\Omega g(x)\,\mathrm{d}\mu)$$

Since μ is a measure, it holds that $\mu(\Omega) = 1$. Therefore $\int_\Omega 1\,\mathrm{d}\mu = 1$. Following the above reasoning, the proof of Jensen's inequality if conducted in HOL4 by dividing the main goal to a number of key lemmas. The first lemma shows the monotonicity of the sub-derivatives:

Theorem 10 (Monotone Sub-derivatives).

$\vdash \forall$(f:real→real) a b x x0 z z0. x < z ∧ z < z0 ∧ z0 < x0 ⇒
$\frac{f(x)-f(z)}{x-z} \leq \frac{f(x0)-f(z0)}{x0-z0} \wedge \lim_{x\to z} \frac{f(x)-f(z)}{x-z} = a \wedge \lim_{x0\to z0} \frac{f(x0)-f(z0)}{x0-z0} = b \Rightarrow$ a≤b

The next theorem uses the monotonicity of the sub-derivatives and the Mean value theorem to prove the existence of the constants a and b provided that f is convex. This gives the inequality $f(x) \geq a.x + b$:

Theorem 11 (Existence of *a* and *b*: *f(x)* ≤ *a.x + b*).

```
⊢ ∀(f:real→real) x x0 z1.(x<z1 ∧ z1<x0 ∧ convex f ∧ f differentiable x0 ∧
    (∀(z2:real).  x < z2 ∧ z2 < x0 ⇒ f differentiable z2) ∧
    (∀z3.  x ≤ z3 ∧ z3 ≤ x0 ⇒ f contl z3) ∧
    (∃l z.  x<z ∧ z<x0 ∧ ((f diffl l) z) ∧ (f(x0)-f(x) = (x0-x)*l))
  ⇒
   ∃f0.  (((f diffl f0) x0) ∧ ((f(x0)-f(x)) ≤ (x0-x)*f0)))
```

where `diffl` is the HOL4 definition of first derivative.

Finally, combining the results of Theorems 10 and 11, we formalize Jensen's inequality for the continuous case:

Theorem 12 (Jensen's inequality).

```
∀f g m.  measure_space m ∧ integrable m g ∧
    (Normal b = integral m λy.(Normal b)) ∧
  ((Normal a)*(integral m λx.g(x)) + (Normal b) = f((integral m λx.g(x)))) ∧
   (∀x.(Normal a)*x + (Normal b) ≤ f(x:extreal))
⇒
  f (integral m λx.g(x)) ≤ integral m λx.(f(g(x)))
```

Since the integral that we use is a Lebesgue integral which is defined over the extended reals, we have to convert our objects to the `extreal` type. This is done by the function `Normal` [13].

6 Conclusion

In this report, we presented the formalizations of the Data Processing Inequality for the discrete case and Jensen's inequality for the continuous case. The former is built on top of the formalizations of the Discrete Time Markov Chains and the probability and information theory. The key lemma of the proof is to show that given a Markov chain $X \to Y \to Z$, for the conditional mutual information it holds that $I(X;Z|Y) = 0$. It is a big challenge to prove the same statement for the continuous case. The difficulty comes from the fact that the definition of the Kullback-Leibler Divergence is very complex. It relies on the definition of the Lebesgue integral over the extended reals. Since the integral is defined as a limit which is itself complex, all related properties should be proved in a way that all definitions are unfolded. This eventually makes the goal extremely complex having the size of several pages. The proof of Jensen's inequality relies on the key proof of the monotonicity of the first derivative of a convex function. It was proven in its measure theoretic form and the HOL code can be found in [5]. A straightforward application of Jensen's inequality is Gibb's inequality [17], which shows that the Kullback-Leibler divergence of the probability distributions p and q, $D_{KL}(p||q)$, is non-negative.

References

[1] HOL4, https://hol-theorem-prover.org/, 2016.

[2] Probability Theory in HOL4, https://hol-theorem-prover.org/kananaskis-10-helpdocs/help/src-sml/htmlsigs/probabilityTheory.html, 2016.

[3] A. R. Coble. *Anonymity, Information, and Machine-Assisted Proof.* PhD thesis, King's College, University of Cambridge, UK, 2010.

[4] T. M. Cover and J. A. Thomas. *Elements of Information Theory.* Wiley-Interscience, 2006.

[5] G. Helali and C. Dunchev. Towards The Quantitative Analysis of Information Flow in HOL, HOL4 code, http://hvg.ece.concordia.ca/projects/prob-it/gainMinEntropy.php, 2015.

[6] J. Hölzl. *Construction and Stochastic Applications of Measure Spaces in Higher-Order Logic*. PhD thesis, Institut für Informatik, Technische Universität München, Germany, October 2012.

[7] L. Liu. *Formalization of Discrete-time Markov Chains in HOL*. PhD thesis, Department of Electrical and Computer Engineering, Concordia University, Montreal, Quebec, Canada, June 2013.

[8] T. Mhamdi. *Information-Theoretic Analysis using Theorem Proving*. PhD thesis, Department of Electrical and Computer Engineering, Concordia University, Canada, December 2012.

[9] T. Mhamdi, O. Hasan, and S. Tahar. On the Formalization of the Lebesgue Integration Theory in HOL. In *Interactive Theorem Proving*, pages 387–402, 2010.

[10] 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.

[11] T. Mhamdi, O. Hasan, and S. Tahar. Quantitative Analysis of Information Flow Using Theorem Proving. In *Formal Methods and Software Engineering*, volume 7635 of *LNCS*, pages 119–134. Springer, 2012.

[12] G. Smith. On the Foundations of Quantitative Information Flow. In *International Conference on Foundations of Software Science and Computational Structures*, pages 288–302. Springer, 2009.

[13] 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, Montreal, Quebec, Canada, January 2011.

[14] T. Mhamdi, O. Hasan and S. Tahar. Formalization of Measure and Lebesgue Integration for Probabilistic Analysis in HOL. In *ACM Transactions on Embedded Computing Systems*, volume 12(1), pages 1 – 23. 2013.

[15] T. Mhamdi, O. Hasan and S. Tahar. Evaluation of Anonymity and Confidentiality Protocols using Theorem Proving. In *Formal Methods in System Design*, volume 47(3), pages 265 – 286. Springer, 2015.

[16] T. Verhoeff. The Laws of Large Numbers Compared, http://www.dklevine.com/archive/strong-law.pdf, 1993.

[17] J. D. Weeks. External fields, density functionals, and the gibbs inequality. In *Journal of Statistical Physics*, volume 110, pages 1209 – 1218. Springer, 2003.