Towards an Axiomatization for the Generalization of the Kullback-Leibler Divergence to Belief Functions

Hélène Soubaras

Thales TRT France, Campus Polytechnique, 1 av. Angustin Fresnel, F–91767 Palaiseau Cedex

Abstract

In his information theory, Shannon [1] defined a notion of uncertainty, the *entropy*, which has been generalized in several wways to belief functions [2]. He also defined the *channel capacity* for which we propose in this paper the first generalization to belief functions. To do that, we need first to generalize the Kullback-Leibler (KL) divergence, for which the present work proposes some axioms. Their list is still not exhaustive since the proposed solution is not unique. But there are many practical interests, since the notion of channel capacity is useful to characterize and optimize for example systems of sensors; its generalization to belief functions allows us to include imprecise sensors such as the human. Finally we show an example of gradient algorithm to compute the generalized channel capacity.

Keywords: Dempster-Shafer theory of belief functions, channel capacity, Kullback-Leibler divergence

1. Introduction

Shannon's fundamental works on information theory [1] introduced the entropy of a random variable, which is a transposition of the entropy in thermodynamics expressed for random particles in physical statistics. The entropy measures the unpredictability. Shannon also defined the mutual information between two random variables, and the channel capacity as the maximal mean mutual information between its input and its output. The channel capacity is the amount of information (in bits) that it is possible to vehicle through a channel. This was the basis for most coding techniques.

A channel can be any physical system providing a measure (the channel output) that is a function of an unknown random variable (the channel input) (see Figure 1). This can be the case for example for any information system consisting of a series of sensors (including human observers) and the associated data fusion [3]. Such information systems are usually evaluated with classical measures such as the probability of error, the false alarm probability and the probability of detection, confusion matrices, or channel capacity. But approaches for their evaluation as a channel capacity has never been proposed in the non-probabilistic case. In classical information theory, the channel input and output are probabilistic random variables.

Here, we consider the case of uncertain systems whose output is not known through its probabilities, but through a mass function as defined in the Demspter-Shafer framework [4]. Measures of uncertainty have been proposed to generalize the entropy to belief functions [2, 5, 6]. But none of these generalizations of information theory dealt with the notion of channel. This paper offers the first mathematical expression that generalizes the measure of channel capacity to channels whose input is a random variable and whose output is a belief function. The proposed axiomtic has been constructred around the pignistic probability of a belief mass since they coincide in the boundary case.

Figure 1: A transmission channel is a way to represent the dependence between an unknown variable (the input X) and an observed one (the output Y).

2. Preliminary

2.1. The KL divergence

In this paper, we consider the notion of Kullback-Leibler (KL) divergence [7, 8] which plays a role in the definition of the capacity of a channel. The channel possible inputs are possible hypotheses that influence the probability distribution of the observations at the channel output. The KL divergence provides a means of comparison of these possible distributions. Its classical expression is as follows

Definition Let p_1 and p_2 be two probability distributions on a same finite space Ω. The Kullback-Leibler (KL) divergence [7, 8] of p_1 w.r.t. p_2 is defined by:

$$D_K(p_1\|p_2) = \sum_{x\in\Omega} p_1(x) \log \frac{p_1(x)}{p_2(x)}$$

It is also called *relative entropy*, or sometimes, improperly, *Kullback distance*.

Indeed it is not a distance. It has the properties:

- it is positive,
- it is zero if and only if $p_1 = p_2$,
- but it is not symmetric: $D_K(p_1\|p_2) \neq D_K(p_2\|p_1)$,
- the triangle inequality $D_K(p_1\|p_3) \leq D_K(p_1\|p_2) + D_K(p_2\|p_3)$ is not satisfied generally.

The quantity $\log \frac{p_1(x)}{p_2(x)}$ is defined as an information in x for discrimination between the two hypotheses 1 and 2. So the KL divergence is the mean information for discriminating between two hypotheses per observation for p_1. Suppose hypothesis 1 is one given input to the channel and hypothesis 2 is "we do not know what is the channel input". If the KL divergence is high, one concludes that the given input provides a quite deterministic observation (so not too uncertain).

The present work proposes a measure for generalizing the mutual information to belief functions. This will allow us to define a channel capacity when the input is a probabilistic random variable but the output is a belief mass. Such channels will be called *uncertain channels*.

2.2. Existing axiomatics

In the literature one can find axiomatic characterizations proposed for measures of dissimilarity that are distances [9, 10]. They lead to 3 axioms: positivity, symmetry w.r.t. the probability densities, and the triangle inequality. Axiomatic characterizations have also been proposed for the generalization of Shannon entropy [11]. An axiomatic approach for the entropy of capacities like Marichal's one was proposed by Kojadinovic [12]. None of them concerns exactly our purpose but close notions; we retained the axioms that could be applied to the KL divergence, and finally, we propose the list of axioms provided in this paper.

2.3. Basics of belief functions

One calls *frame of discernment* a set Ω of all possible mutually exclusive hypotheses; it will be supposed finite.

A *mass function* [4], is a set function from 2^Ω to $[0; 1]$ such that $\sum_A m(A) = 1$. A subset $A \subseteq \Omega$ is called a *focal set* as soon as its mass is non zero. Let $\mathcal{F}$ denote the set of focal sets. m becomes a classical probability when the focal sets are disjoint singletons. It is then said *Bayesian*. In the sequel we will suppose that $m(\emptyset)$ is null; i.e. the mas is normalized.

The belief function *Bel* and the plausibility function *Pl* are defined from the mass as follows:

$$Bel(A) = \sum_{B \subseteq A} m(B) \text{ and } Pl(A) = \sum_{B \cap A \neq \emptyset} m(B)$$

And the ignorance is $Ig(A) = Pl(A) - Bel(A)$. The *pignistic probability* [13] *Bet* is defined for all $x \in \Omega$ by:

$$Bet(x) = \frac{1}{1 - m(\emptyset)} \sum_{A/x \in A} \frac{m(A)}{|A|}$$

where $|A|$ is the cardinality of A, i.e. the number of elements of A.

We will define, for any mass function, an associated pignistic mass as follows. It will be useful for its particular properties while studying the generalization of the KL divergence.

Definition Let m be a belief mass on a frame Ω and let $\mathcal{F}$ be its set of focal sets. The *pignistic mass* associated to m is the mass *mBet* on the same frame Ω whose sets of focal sets $\mathcal{F}_{Bet}$ is the partition generated by $\mathcal{F}$ (i.e. the smallest partition such that each focal set is a union of sets of $\mathcal{F}_{Bet}$), and whose mass values are the pignistic probabilities of m, i.e., for all $A_i \in \mathcal{F}_{Bet}$:

$$m_{Bet}(A_i) = Bet(A_i) = \sum_{x \in A_i} Bet(x) = \sum_{B \in \mathcal{F}} m(B) \frac{|A_i \cap B|}{|B|}$$

Note that m_{Bet} is a probability. It is Bayesian and equal to *Bet* if the sets A_i are all singletons.

3. Axioms and properties for a generalization of the KL divergence

Let Ω be a frame of discernment. Let m_1 and m_2 be two mass functions on Ω. $\mathcal{F}_1$ and $\mathcal{F}_2$ are respectively the sets of focal sets of m_1 and m_2. They are supposed to be finite. We denote as $\overline{D_K}(m_1\|m_2)$ the generalized KL divergence of m_2 in comparison to m_1. Our objective is to propose a mathematical expression for it. In order to make sense, there are some properties that must be verified. Their list is proposed in this section.

We will suppose that the considered uncertain channel has a finite number N_x of possible inputs x and a finite number N_f of focal sets A_i, $1 \leq i \leq N_f$ at its output. The marginal belief mass m of any focal set A is:

$$m(A) = \sum_x m_x(A) p(x) = \mathbb{E}_X[m_x(A)] \tag{1}$$

The notion of *independence* is important in our purpose since it characterizes the fact that there is no relation between two variables. Some definitions of independence between two belief masses have been proposed [14, 15]. But here as we consider a belief mass and a probability we will keep a probabilistic definition. It corresponds to the particular case of *irrelevance* [14] between two belief masses: "knowing X does not affect belief on Y". So, the input X and the output Y of the channel will be said independent if and only if, for all couple (x, A), $m_x(A) = m(A)$.

3.1. Axioms: boundary condition, positiveness and symmetry

The expression must be compatible with the probabilistic case. When Ω is finite, if one has two probability distributions p_1 and p_2 on Ω, the classical KL divergence is [7]:

$$D_K(p_1\|p_2) = \sum_{x \in \Omega} p_1(x) log \frac{p_1(x)}{p_2(x)} \tag{2}$$

$\overline{D_K}$ must coincide with D_k when m_1 and m_2 are pure probabilities and then said *Bayesian*, i.e. their focal sets are the singletons of Ω.

Positiveness: $\overline{D_K}(m_1||m_2) \geq 0$ must always be satisfied [9].

Symmetry: It must be independent of the ordering of Ω elements [9].

3.2. Property: extended boundary condition

We extend the boundary condition for additive measures where the focal sets are a partition. This is a restricted probabilistic case. If Ω is finite or if it is not, we can extend this property as follows: if $\mathcal{F}_1$ is a partition $\{A_1, A_2, ...A_{Nf}\}$ of Ω, and if $\mathcal{F}_2 = \mathcal{F}_1$, we must have a generalized entropy of the form

$$\overline{D_K}(m_1||m_2) = D_K(m_1||m_2) = \sum_i m_1(A_i) \log \frac{m_1(A_i)}{m_2(A_i)} \tag{3}$$

3.3. Property: decreases when replacing a mass by its pignistic mass

When one of the two masses m_1 or m_2 is replaced by its corresponding pignistic mass, $\overline{D_K}$ must decrease. This expresses the intuitive notion that when a mass becomes a probability, if has less ignorance and less imprecision.

3.4. Axiom: approximately minimal for the pignistic mass of m_1

We propose here a definition to characterize the fact that two values are approximately equal. The more their associated belief masses will be far from the bayesian case, the more their equality will be inaccurate. We will say that the values are equal *up to an uncertainty*.

When m_1 is fixed, $\overline{D_K}(m_1||m_2)$ must be minimum, in a way that will be called *up to an uncertainty*, when $m_2 = mBet_1$ (the pignistic mass of m_1). That means there exists a function $u(m_1)$ which is zero when m_1 is Bayesian and such that

$$\overline{D_K}(m_1||m_2) - u(m_1)$$

is minimum for $m_2 = mBet_1$

This axiom is proposed to express the intuitive idea that the generalized KL divergence is equal to the KL divergence of probabilities that are "close" to the masses m_1 and m_2 (null when these probabilities are equal), plus an uncertainty term which is related to the "distance" of these masses to these probabilities.

3.5. Property: divergence to the total ignorance is (almost) the uncertainty

Whe know that Shannon entropy H and the classical definition of the KL divergence for a probability p, when p_0 denotes the uniform law on Ω, satisfy the property:

$$D_K(p||p_0) = \log(|\Omega|) - H(p)$$

Let m_0 denote the mass function of the total ignorance. That means it has one single focal set which is Ω. There must exist a measure of uncertainty U satisfying for each mass m:

$$|\overline{D_K}(m||m_0) - \varphi(m, |\Omega|) + U(m)| \leq u(m)$$

where φ and u are functions to determine. U must coincide with the entropy and u must be zero in the purely probabilistic case.

We introduce U to represent the generalized entropy associated to the generalized KL divergence, because we will use it to express several properties. The auxiliary function u is a tool to express the fact that a property is "almost" true up to an uncertainty due to the fact that a belief mass is not a true probability.

3.6. Associated mutual information

Remind that the basic idea it to measure the capacity of an uncertain channel. At its output (the unknown random variable known through m), Y has the marginal mass (given in Equation 1).

Shannon's definition of the mutual information $I(X;Y)$ [1] can be generalized to belief functions as $\bar{I}(X;Y)$ by: it exists a function $u(m)$, which becomes zero in the probabilistic case, such that

$$|\bar{I}(X;Y) - \mathbb{E}_X[\overline{D_K}(m_x||m)]| \leq u(m)$$

Knowing that in the probabilistic case, one has also the following equality for the mutual information:

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

where H is Shannon entropy, one would like to have in the generalized case. This equality will be true *more or less a belief mass uncertainty*: i.e. there exists a function $u(m)$ which s zero if m is Bayesian and satisfying

$$|\mathbb{E}_X[\overline{D_K}(m_x||m)] - U(m) + \mathbb{E}_X[U(m_x)]| \leq u(m) \tag{4}$$

where U is the measure of uncertainty of Section 3.5.

3.7. Axiom: the independent case

Let us consider again the channel which input is a random variable X and which output is an belief mass on Y. Let m denote the marginal mass for Y. X and Y will be independent in a generalized meaning if and only if the generalized mutual information is zero *more or less a belief mass uncertainty*, i.e. there exists a function $u(m)$ which is zero if m is Bayesian and such that

$$\bar{I}(X;Y) \leq u(m)$$

Furthermore, the uncertainty U associated to $\overline{D_K}$ must satisfy in that case that there exists also a function $v(m)$ which is zero in the Bayesian case and such that

$$|U(X,Y) - H(X) - U(Y)| \leq v(m)$$

where U is the measure of uncertainty of Section 3.5.

4. The new proposed generalization

4.1. Idea

Let m_1 and m_2 be two masses on a common frame Ω. Their sets of focal sets are $\mathcal{F}_1$ and $\mathcal{F}_2$ respectively; they are not partitions in the general case. One must first find a partition of all the possible events. Furthermore, our idea is to exploit the *Bel* and the *Pl* functions as bounds of the probabilities of the subsets. So one needs a partition into subsets where the pignistic probability is a constant. So the best fitted partition $\mathcal{H}$ is the partition generated by $\mathcal{F}_1 \cup \mathcal{F}_2$.

We are going to express the KL divergence for two of the probability densities p_1 and p_2 that are compatible with the mass functions, and then to propose an approximation/majoration that takes into account the ignorance Ig. Let us consider one set $A_i \in \mathcal{H}$. The ignorance $Ig(A_i)$ is the length of interval for $p(A_i)$:

$$Bel(A_i) \leq p(A_i) = \sum_{x \in A_i} p(x) \leq Pl(A_i)$$

Furthermore, one can write for all $x \in A_i$

$$p(x) = \frac{Bet(A_i)}{|A_i|} + \varepsilon(x) \tag{5}$$

with for all x, $|\varepsilon(x)| \leq 1$ and

$$|\sum_{x \in A_i} \varepsilon(x)| \leq Ig(A_i)$$

Proof. This is true because

$$0 \leq p(A_i) - Bel(A_i) = \sum_{x \in A_i} p(x) - Bel(A_i)$$

$$= Bet(A_i) - Bel(A_i) + \sum_{x \in A_i} \varepsilon(x) \leq Ig(A_i)$$

and

$$-Ig(A_i) \leq Bel(A_i) - Bet(A_i) \leq 0$$

since $Bel(A_i) \leq Bet(A_i) \leq Pl(A_i)$. So, by adding the two double inequalities:

$$-Ig(A_i) \leq \sum_{x \in A_i} \varepsilon(x) \leq Ig(A_i)$$

■

The quantity $\varepsilon(x)$ represents the difference between one of the possible probabilities p and the probability which is uniform inside the set A_i, which is the pignistic probability *Bet*. Thus we introduced the above expression for $p(x)$ in Equation 5 in order to express further a limited development for the KL divergence.

Thus, for p_1 and p_2, the classical KL divergence D_K should be of the form

$$D_K(p_1 \| p_2) = -H(p_1) - \sum_i \sum_{x \in A_i} p_1(x) \log p_2(x)$$

$$= -H(p_1) + \alpha$$

So, α is equal to

$$-\sum_i \sum_{x \in A_i} \left(\frac{Bet_1(A_i)}{|A_i|} + \varepsilon_1(x) \right) \times \log \left(\frac{Bet_2(A_i)}{|A_i|} + \varepsilon_2(x) \right)$$

4.2. Introducing a limited development

If one considers the order 1 limited development of $D_K(p_1 \| p_2)$, note that $\log(1 + y) \sim y$ for y close to 0 so the log term is equivalent to

$$\log \left(\frac{Bet_2(A_i)}{|A_i|} + \varepsilon_2(x) \right)$$

$$\sim \log \frac{Bet_2(A_i)}{|A_i|} + \frac{|A_i|}{Bet_2(A_i)} \varepsilon_2(x)$$

This approximation is true when ε_2 is small. It will help us to find out an expression for an upper bound of D_k in this case. When ε_2 is no longer small (this may occur !), we do not prove that it is still an upper bound; but this is not our purpose. The idea is to verify that this expression satisfies the required axioms and properties for the generalization $\overline{D_K}$.

Then α is equivalent (at order 1) to

$$\beta = -\sum_i \sum_{x \in A_i} \left(\frac{Bet_1(A_i)}{|A_i|} + \varepsilon_1(x) \right)$$

$$\times \left(\log \frac{Bet_2(A_i)}{|A_i|} + \frac{|A_i|}{Bet_2(A_i)} \varepsilon_2(x) \right)$$

The above expression of β is equivalent at order 1 to

$$\gamma = -\sum_i Bet_1(A_i) \log \frac{Bet_2(A_i)}{|A_i|}$$

$$-\sum_i \log \frac{Bet_2(A_i)}{|A_i|} \sum_{x \in A_i} \varepsilon_1(x) - \sum_i \frac{Bet_1(A_i)}{|A_i|} \frac{|A_i|}{Bet_2(A_i)} \sum_{x \in A_i} \varepsilon_2(x)$$

by neglecting the term containing $\varepsilon_1(x)\varepsilon_2(x)$. This term is bounded upperly by the quantity δ:

$$\delta = -\sum_i Bet_1(A_i) \log \frac{Bet_2(A_i)}{|A_i|}$$

$$-\sum_i Ig_1(A_i) \log \frac{Bet_2(A_i)}{|A_i|} + \sum_i \frac{Bet_1(A_i)}{Bet_2(A_i)} Ig_2(A_i)$$

Proof. This is true because if one has for all i, $a_i \geq 0$ and $|b_i| \leq B_0$ then the following inequality is true:

$$\sum_i a_i b_i \leq B_0 \sum_i a_i$$

since $-a_i B_0 \leq a_i b_i \leq a_i B_0$ for all i. Precisely, for all i, we have:

$$-\log \frac{Bet_2(A_i)}{|A_i|} \geq 0 \text{ and } \frac{Bet_1(A_i)}{Bet_2(A_i)} \geq 0$$

■

So we propose the formula given next Section.

4.3. The proposed expression

Let Ω be a frame, and let m_1 and m_2 be two mass functions on Ω. Their sets of focal sets are respectively $\mathcal{F}_1$ and $\mathcal{F}_2$. Let $A_i \subset \Omega$ denote the subsets of the partition generated by $\mathcal{F}_1 \cup \mathcal{F}_2$.

The proposed generalization of KL divergence to belief functions of m_2 in comparison to m_1 is

$$\overline{D_K}(m_1\|m_2) = -U(m_1) - \sum_i (Bet_1(A_i) + Ig_1(A_i)) \log \frac{Bet_2(A_i)}{|A_i|} + \sum_i \frac{Bet_1(A_i)}{Bet_2(A_i)} Ig_2(A_i) \quad (6)$$

where U is a measure of uncertainty, generalizing Shannon entropy. It must satisfy $U(m) \leq 0$, and be zero in the deterministic case.

4.4. Expression of the uncertainty U

We are going to show that the satisfaction of Property 3.6 in Equation 4 leads to an expression for the above generalized uncertainty U.

Remember that we consider a discrete random variable X (the channel input) with probabilities $p(x)$, and a mass function m_x depending on X (the channel output). As all the functions $F = m$, Bel, Pl, Bet and Ig are linear w.r.t. the mass m_x, their average over X is:

$$F = \mathbb{E}_X[F_x] = \mathbb{E}_X[F(m_x)] = \sum_x F(m_x)p(x)$$

So to satisfy the property of the generalized mutual information (given in Section 3.6) we can write, using Equation 6:

$$\overline{D_K}(m_x\|m) = -U(m_x) - \sum_i (Bet_x(A_i) + Ig_x(A_i)) \log \frac{Bet(A_i)}{|A_i|} + \sum_i Ig(A_i) \frac{Bet_x(A_i)}{Bet(A_i)}$$

So

$$\mathbb{E}_X[\overline{D_K}(m_x\|m)] - U(m) + \mathbb{E}_X[U(m_x)] = -U(m) + \sum_i Ig(A_i) - \sum_i (Bet(A_i) + Ig(A_i)) \log \frac{Bet(A_i)}{|A_i|}$$

The property is satisfied if the following sufficient condition is verified:

$$U(m) = -\sum_i Bet(A_i) \log \frac{Bet(A_i)}{|A_i|} \quad (7)$$

This is the entropy of the pignistic probability $H(Bet)$. This expression of U obviously coincides with the classical entropy when the mass is Bayesian. So the resulting expression for the generalized KL divergence is

$$\overline{D_K}(m_1\|m_2) = \sum_i Bet_1(A_i) \log \frac{Bet_1(A_i)}{Bet_2(A_i)} - \sum_i Ig_1(A_i) \log \frac{Bet_2(A_i)}{|A_i|} + \sum_i \frac{Bet_1(A_i)}{Bet_2(A_i)} Ig_2(A_i) \quad (8)$$

One can also write that it is equal to

$$D_K(Bet_1\|Bet_2) + \sum_i Ig_1(A_i) \log \frac{|A_i|}{Bet_2(A_i)} + \sum_i Ig_2(A_i) \frac{Bet_1(A_i)}{Bet_2(A_i)} \quad (9)$$

The first term represents the KL divergence of the pignistic probabilities, which would be the most "reasonable" choices to replace the masses by probabilities. The second term shows that the contribution of Bet_2 to the uncertainty relatively to one probability of the core of m_1 is enlarged by the probability interval of m_1. The third term is the direct incidence of the probability interval of m_2 corrected by the fact that it is considered relatively to Bet_1. In other words, the second and the third term show the incidence of the probability intervals in m_1 and in m_2 respectively.

Note that we proposed a sufficient condition without proving whether it is necessary or not. The solution may be non unique.

5. Verification of the properties

5.1. Axiom 3.1: boundary condition, positivenesss and symmetry

If m_1 and m_2 are Bayesian, the sets A_i are the singletons of Ω. Their ignorance is zero. Thus, $\overline{D_K}$ becomes

$$\overline{D_K}(m_1\|m_2) = -H(m_1) - \sum_i m_1(A_i) \log m_2(A_i)$$

This is the classical KL divergence.

The positiveness is obviously satisfied, since if we consider the expression of $\overline{D_K}$ in Equation 9, there are three terms. We know that the classical KL divergence, which is the first term, is positive, and all the terms of the following sums are positive since the following inequality is always true:

$$\log \frac{|A_i|}{Bet_2(A_i)} \leq 0$$

The symmetry is also obviously satisfied since the ordering of Ω does not intervene in the definition of the sets A_i.

5.2. Property 3.2: extended boundary condition

If $\mathcal{F}_1 = \mathcal{F}_2$ is a finite partition, then we have $Bet_1(A_i) = m_1(A_i)$, $Bet_2(A_i) = m_2(A_i)$, $Ig_1(A_i) = 0$ and $Ig_2(A_i) = 0$ for all i. Thus the expression 9 becomes

$$\overline{D_K}(m_1\|m_2) = D_K(m_1\|m_2)$$

So the property is satisfied.

5.3. Property 3.3: decreases when replacing m by the pignistic mass

This is true since Ig_1 or Ig_2 are positive and they become zero in those cases.

5.4. Axiom 3.4: approximately minimal for the pignistic mass

Let us define

$$u(m_1) = \sum_i Ig_1(A_i) \log \frac{|A_i|}{Bet_2(A_i)}$$

It is zero if m_1 is Bayesian. Thus

$$\overline{D_K}(m_1 \| m_2) - u(m_1)$$

$$= D_K(Bet_1 \| Bet_2) + \sum_i Ig_2(A_i) \frac{Bet_1(A_i)}{Bet_2(A_i)}$$

The second term is minimal and equal to zero if and only if the ignorance of m_2 is zero for all A_i. And one knows that the first term, which is a classical KL divergence, is minimal and zero for $Bet_2 = Bet_1$. Thus the minimal is zero and it is reached when $m_2(A_i) = Bet_1(A_i)$ for each subset A_i. The property is satisfied.

5.5. Axiom 3.6: mutual information

As seen in Section 4.4, this property is satisfied by definition of U. The generalized mutual information is then

$$\overline{I}(X;Y) = \mathbb{E}_X[\overline{D_K}(m_x \| m)] = \mathbb{E}_X[A + B + C] \qquad (10)$$

where

$$A = \sum_i Bet_x(A_i) \log \frac{Bet_x(A_i)}{Bet(A_i)}$$

$$B = \sum_i Ig_x(A_i) \log \frac{|A_i|}{Bet(A_i)} \sum_i Ig_x(A_i) \log \frac{|A_i|}{Bet(A_i)}$$

$$C = \sum_i Ig(A_i) \frac{Bet_x(A_i)}{Bet(A_i)}$$

$$= \mathbb{E}_X \left[\sum_i Bet_x(A_i) \log Bet_x(A_i) \right] - \sum_i Bet(A_i) \log Bet(A_i)$$

$$+ \sum_i Ig(A_i) \left(1 + \log \frac{|A_i|}{Bet(A_i)} \right)$$

This can also be written as

$$\overline{I}(X;Y) = H(Bet(Y)) - H(Bet(Y|X))$$

$$+ \sum_i Ig(A_i) \left(1 + \log \frac{|A_i|}{Bet(A_i)} \right) \qquad (11)$$

5.6. Axiom 3.7: the independent case

We find, using the expression of the mutual information in Equation 11, in the independent case:

$$\overline{I}(X;Y) = \sum_i Ig(A_i) \left(1 + \log \frac{|A_i|}{Bet(A_i)} \right)$$

5.7. Property 3.5: divergence to the total ignorance is (almost) the uncertainty

In that case the sets A_i are the partition generated by $\mathcal{F}_1$. For all i, we have $Bet_2(A_i) = \frac{|A_i|}{|\Omega|}$ and $Ig_2(A_i) = 1$. So

$$\overline{D_K}(m_1 \| m_0) = \sum_i Bet_1(A_i) \log \frac{Bet_1(A_i)|\Omega|}{|A_i|}$$

$$+ \sum_i Ig_1(A_i) \log |\Omega| + \sum_i \frac{Bet_1(A_i)|\Omega|}{|A_i|}$$

$$= -U(m_1) + \log |\Omega| \left(1 + \sum_i Ig_1(A_i) \right) + |\Omega| \sum_i \frac{Bet_1(A_i)}{[A_i|}$$

Thus the property is satisfied.

6. Generalized channel capacity

We have proposed a measure for generalizing the mutual information to belief functions. This will allow us to define a channel capacity for DS output-modeled channels, called *uncertain channels*, and to propose a methodology for evaluation and optimization of such systems.

6.1. The new channel capacity

The generalization of the mutual information between the input and the output of an uncertain channel has been proposed in Section 5.5 (Equation 5.5). It will be expressed here in bits, so we take the 2-basis logarithm. We are going to define now some matrices to express this mutual information. There exist a $(N_f \times N_f)$-sized matrix N which computes the vector Bet of the pignistic probabilities of the focal sets as a function of the mass vector and a matrix A to compute the vector of the ignorance of the focal sets:

$$Bet = N.M \text{ and } Ig = A.M$$

Let us also introduce a third matrix B of coefficients

$$b_{ij} = \frac{n_{ij}}{|A_i|}$$

and the N_f-sized vector I_1 whose coordinates are all equal to 1. Finally, let us define a function λ which maps the positive elements of a matrix into their 2-basis logarithm, and the other elements to 0. Thus, the mutual information can be expressed as a matrix product as follows:

$$\overline{I}(X;Y) = \mathbb{E}_X \left[(N.M_x)^T . \lambda(N.M_x) \right] - (N.M)^T . \lambda(N.M)$$

$$+ (A.M)^T . (I_1 - \lambda(B.M)) \qquad (12)$$

As M is the marginal mass vector resulting from all the mass vectors M_x for each input x of the channel, one can introduce the matrix K_m such that:

$$M = \mathbb{E}_X [M_x] = K_m.P$$

So $K_m(i,j) = m_{x_j}(A_i)$.

We have expressed the mutual information as a function of the vector P of the probabilities p_x. To obtain the channel capacity, one just has to maximize this mutual information over all possible input probability densities P. This can be performed for example with a gradient method.

6.2. Example: optimization with the gradient algorithm

Note that for any $n \times n$ matrices F and G and the n-sized vector $X = (x_1, x_2...x_n)^T$, if $G.X$ coordinates are > 0:

$$\frac{\partial}{\partial x_i}(F.X)^T \lambda(G.X) = \frac{\partial}{\partial x_i} \sum_j \left(\sum_k f_{jk} x_k \right) log_2 \left(\sum_l g_{jl} x_l \right)$$

$$= \sum_j f_{ji} log_2 \left(\sum_l g_{jl} x_l \right) + \sum_j g_{ji} \frac{\sum_k f_{jk} x_k}{Log(2) \sum_l g_{jl} x_l}$$

The gradient of the mutual information (Equation 12) as a function of the probability vector P of coordinates $p_i = Pr(X = x_i)$ is then

$$\frac{\partial \bar{I}(X;Y)}{\partial p_i} =$$

$$(N.M_{x_i})^T.\lambda(N.M_{x_i}) - \frac{1}{Log(2)} \left(\sum_j h_{ji} (1 + Log \left(\sum_l h_{jl} p_l \right) \right)$$

$$+ \sum_j f_{ji} - \sum_j f_{ji} log_2 \left(\sum_l g_{jl} p_l \right) + \sum_j g_{ji} \frac{\sum_k f_{jk} p_k}{Log(2) \sum_l g_{jl} p_l}$$

with $F = A.K_m$ and $G = B.K_m$ and $H = N.K_m$. And thus the gradient is the N_x-sized vector $Grad_I$ whose coordinates are $\frac{\partial \bar{I}(X;Y)}{\partial p_x}$. It can be expressed as:

$$Grad_I = Grad_0 - \frac{1}{Log(2)} H^T.I_1 - H^T.\lambda(H.P) + F^T.I_1$$

$$-F^T.\lambda(G.P) - \frac{1}{Log(2)} G^T.(F.P \oslash G.P)$$

where $Grad_0$ is the vector of coordinates $(N.M_{x_i})^T.\lambda(N.M_{x_i})$ and $\oslash$ denotes the element-wise division of matrices.

One must perform the optimization under the constraint that P satisfies $\sum_i p_i = 1$, so one must remain orthogonal to the N_x-sized vector $U = (11...1)^T$.

To update the vector P iteratively to optimize I remaining in the domain that orthogonal to U, one can do:

$$P(n+1) = P(n) + \varepsilon(Grad_I(n) - U^T.Grad_I(n).U)$$

where ε is a coefficient inferior or equal to a fixed value (chosen to maintain P coordinates > 0). The algorithm is initialized with a uniform law for P. It is stopped when $P(n+1) - P(n)$ is inferior to a threshold.

7. Application: method for information systems evaluation

Let us consider a system made of sensors (that can also be human) which provides (with possible data fusion) an information modeled by a belief mass. To compute its channel capacity, one needs to measure the parameters $m(A|x)$ for all the focal sets A and all the possible inputs x. We describe here one possible method:

7.1. Measurements

7.1.1. Tests

One just needs to put successively all the possible values for x at the system input, and the values $m(A|x)$ are directly measured in one single test for each x.

This exploits one advantage of belief functions in comparison to probabilistic approaches: they provide directly (in a single time step) a belief mass, while probabilities are not measured directly but after averaging on a number of time steps.

7.1.2. Test and average

When the belief mass obtained for one given input is not totally deterministic (i.e. it may be noisy), the idea is to make several measures of belief masses for one given input x and to average them in order to get a more meaningful value. If the number of trials is N, the overall estimated value of the mass is then, for one focal set A:

$$\hat{m}(A|x) = \sum_n m_n(A|x)$$

where $m_n(A|x)$ is the mass obtained for set A at the n^{th} trial.

7.1.3. Using extra source of knowledge

When it is not possible to choose the channel input x, one must use an additive system which provides an estimation $\hat{x}$ of the actual input x. This estimation may be provided by another system, by a human expert, or by the system itself (at the end of the signal processing chain). As the estimation $\hat{x}$ may be erroneous, the result will be better by averaging:

$$\hat{m}(A|x) = \sum_n m_n(A|\hat{x} = x)$$

7.2. Optimization

7.2.1. Adjusting the input statistics

Once the parameters $m(A|x)$ are all known, one can compute directly the optimization of the mean mutual information to obtain at the same time the channel capacity and the corresponding input probability distribution P_x.

This is interesting in practice when it is possible to control P_x. This is possible for example, when one of

the sensors is a human, by adjusting the criteria from which he will have to provide a given alarm.

7.2.2. Adjusting the system itself

It is also possible to optimize the mean mutual information by modifying the systems itself. It can be a parameter as for example the location of one sensor, or a fusion method.

8. Conclusion

We proposed a measure for generalizing the mutual information to belief functions. This can allow us to define a channel capacity for belief mass output-modeled channels, and to propose a methodology for evaluation and optimization of such systems.

We started an axiomatization to generalize the KL divergence to belief functions. The axioms are the boundary condition (in the probabilistic case), the positiveness, the symmetry w.r.t. the ordering of the elements, the (almost) minimum for the pignistic probability, and the (almost) nullity of the associated mutual information in the independent case. A list of properties is also given.

A reasoning is presented to obtain one expression for this generalized KL divergence. But we did not prove that this was the unique solution. so the axiomatic is not complete. To achieve it, some axioms may be added in order to guarantee a unique solution for the generalized KL divergence. For example, one could express as an axiom that the resulting capacity must not exceed its maximal probabilistic value $\log N$, where N is the number of possible inputs.

In the field of belief functions, the literature offers a panel of measures for uncertainty, including generalizations of Shannon entropy. But these works concern only one part of the Information Theory. We have proposed to generalize the other part by defining the first extension of Shannon channel capacity to belief functions. It offers then a generalization of the mutual information, and a definition of the KL divergence for mass functions. They are derived from the expression of discord.

This new measure allows us to evaluate information systems, and to increase their performance, either by controlling the input statistics, if this is possible in practice, or by modifying the system itself (design and date fusion algorithms). We showed and example of algorithm to compute the generalized capacity, and a possible application.

References

[1] Claude Shannon. A mathematical theory of communication. *Bell Syst. Technical J.*, 27:379–423 and 623–656, July and Oct. 1948.

[2] George J. Klir. *Uncertainty and Information. Foundations of Generalized Information Theory*. Wiley, US, 2006.

[3] Isabelle Bloch. Defining belief functions using mathematical morphology ; application to image fusion under imprecision. *Int. J. Approximate Reasoning*, 48:437–465, June 2008.

[4] Glenn Shafer. *A Mathematical Theory of Evidence*. Princeton Univ. Press, Princeton, NJ, 1976.

[5] Nikhil R. Pal, James C. Bezdek, and Rohen Hemasinha. Uncertainty measures for evidential reasoning I: A review. *Int. J. Approximate Reasoning*, 7:165–183, 1992.

[6] Simon Petit-Renaud. *Application de la théorie de l'information et des systèmes flous à l'estimation fonctionnelle en présence d'informations incertaines ou imprécises*. Ph.d. thesis, Université de Technologie de Compiègne, France, 1999.

[7] Solomon Kullback and Richard Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22:79–86, 1951.

[8] Richard E. Blahut. *Principles and practice of information theory*. Addison Wesley Publishing Company, 1990.

[9] Frank Critchley, Paul Marriott, and Mark Salmon. Preferred point geometry and the local differential geometry of the Kullback-Leibler divergence. *The Annals of Statistics*, 22:1587–1602, Sept. 1994.

[10] Carlo Bertoluzza, Mario Di Bacco, and Viviana Doldi. An axiomatic characterization of the measures of similarity. *Sankhya, the Indian J. of Statistics*, 66:474–486, Aug. 2004.

[11] Imre Csiszár. Axiomatic characterizations of information measures. *Entropy*, 10:261–273, 2008.

[12] Ivan Kojadinovic, Jean-Luc Marichal, and Marc Roubens. An axiomatic approach to the definition of the entropy of a discrete Choquet capacity. *Inf. Comput. Sci.*, 172(1–2):131–153, 2005.

[13] Philippe Smets and Robert Kennes. The transferable belief model. *Artificial Intelligence*, 66:191–234, 1994.

[14] Boutheima Ben Yaghlane, Philippe Smets, and Khaled Mellouli. Belief function independence: I. the marginal case. *Int. J. Approximate Reasoning*, 29:47–70, Jan. 2002.

[15] Boutheima Ben Yaghlane, Philippe Smets, and Khaled Mellouli. Belief function independence: II. the conditional case. *Int. J. Approximate Reasoning*, 31:31–75, Oct. 2002.