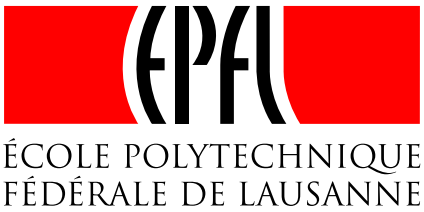

ON THE INEQUALITIES IN INFORMATION THEORY

RETHNAKARAN PULIKKOONATTU

ABSTRACT. Claude Elwood Shannon in 1948, then of the Bell Telephone Laboratories, published one of the most remarkable papers in the history of engineering [1]. This paper ("A Mathematical Theory of Communication", Bell System Tech. Journal, Vol. 27, July and October 1948, pp. 379 - 423 and pp. 623 - 656) laid the groundwork of an entirely new scientific discipline, *information Theory*, that enabled engineers for the first time to deal quantitatively with the elusive concept of *information*".

In his celebrated work, Shannon nicely laid the foundation for transmission and storage of information. Using a probabilistic model, his Theory helped to get further insight into what is achievable and what is not, in terms of quantifiable information transfer. Indeed the very same concept is used to predict the limits on data compression and achievable transmission rate on a probabilistic channel.These underlying concepts can be thought of as inequalities involving measures of probability distributions. Shannon defined several such basic measures in his original work. The field of Information Theory grew with researchers finding more results and insights into the fundamental problem of transmission of and storage using probabilistic models. By nature of the subject itself, the results obtained are usually inequalities involving basic Shannon's measures such as entropies. Some of them are elementary, some rather complicated expressions. In order to prove further theorems as well it required to check whether certain expressions are true in an Information Theoretic sense. This motivated researchers to seek a formal method to check all possible inequalities. Raymond Yeung [2] in 1998 came out with a remarkable framework, which could verify many of the inequalities in this field. His framework thus enabled to verify all inequalities, derived from the basic Shannon measure properties.

A central notion of Information Theory is entropy, which Shannon defines as measure of information itself. Given a set of jointly distributed random variables $X_1, X_2, \ldots, X_n$, we can consider entropies of all random variables $H(X_i)$, entropies of all pairs $H(X_i, X_j)$, etc. ($2^n - 1$ entropy values for all nonempty subsets of $\{X_1, X_2, ..., X_n\}$). For every n-tuple of random variables we get a point in $\mathbb{R}^{2^n-1}$, representing entropies of the given distribution. Following [2] we call a point in $\mathbb{R}^{2^n-1}$ constructible if it represents entropy values of some collection of n random variables. The set of all constructible points is denoted by Γ_n^*

It is hard to characterize Γ_n^* for an arbitrary n (for $n \geq 3$, it is not even closed [**?**]). A more feasible (but also highly non-trivial) problem is to describe the closure $\bar{\Gamma}_n^*$ of the set Γ_n^*. The set $\bar{\Gamma}_n^*$ n is a convex cone [**?**], and to characterize it we should describe the class of all linear inequalities of the form

$$\lambda_1 H(X_1) + \ldots + \lambda_n H(X_n) + \lambda_{1,2} H(X_1 X_2) + \ldots +$$
$$\lambda_{1,2,3} H(X_1, X_2, X_3) + \ldots + \lambda_{1,2,3,\ldots,n} H(X_1, X_2, X_3, \ldots, X_n)$$

which are true for any random variables $X_1, X_2, \ldots, X_n$ (λ_i are real coefficients).

Information inequalities are widely used for proving converse coding theorems in Information Theory. Recently interesting applications of information inequalities beyond Information Theory were found [10],[12],[14]. So investigation of the class of all valid information inequalities is an interesting problem. We refer the reader to [15] for a comprehensive treatment of the subject.

Yeung's framework thus helped to verify all the Shannon type inequalities. Yeung and Yan have also developed a software, to computationally verify such inequalities. Since the software is rather outdated, we have made an attempt to make a more efficient and user friendly implementation of the software, hinging from the original work of Yeung. The software, which we call information inequality solver (iis) is freely available for download from EPFL website. The new software suit has the added advantage that it is freed of dependencies on any licensed products such as Matlab (or toolboxes).

Date: 2008 Januay 12.

Key words and phrases. Inequalities in Information theory,Shannon type inequalities, Xitip.

Laboratory of Information and Communication Systems, Information processing Group, EPFL, Lausanne, Switzerland,1005, ipg.epfl.ch.

Acknowledgments

I am most grateful to Suhas Diggavi for providing opportunity to work on this problem, which I consider as a rewarding and satisfying experience. In spite of being extremely busy, at both personal as well as professional front, he showed faith and interest in this project which was motivating for me to make the best out of it. His suggestions and ideas surely made the software much better than I had initially anticipated to be.

My foremost acknowledgment is to Etienne Perron who has been nothing short of amazing a person to work with. He was always available for my questions, and his prior knowledge in this subject made my life a lot easier. In more than one sense this work is as much his as it is mine, if not more. It is he who initiated this work to develope a C based tool, from where we together made a little nicer suit than it originally was.

Emre Telatar deserve a thank you note for the wonderful course on Information theory as well as for providing constant encouragement during this semester project.

Many thanks to my friends Soni PM, Nandakishire Santhi, Prakash R, Vivek Shenoy and Mahesh Daisy who helped with valuable leads to get the software successfully compile in Windows/cygwin platform.

I also wish to thank the several other people who directly or indirectly helped me in successfully completing this work, most notably Christine Neuberg. She was very kind and supportive in holding discussions in their office.

My most important acknowledgment is to Maya who has filled my life with joy and who means the world to me. This has surely been the hardest time I had to be away and that made me work a little extra hard, to make this as good as I could.

CONTENTS

1. Information Theory: Concept of Information

In his seminal work[1], which literally gave birth to the field of Information Theory, Shannon laid the foundation of transmission and storage of information. Using a probabilistic model, his Theory helped to get further insight into what is achievable and what is not, in terms of quantifiable information transfer. Indeed the very same concept is used to establish the limits on data compression and achievable transmission rate on a probabilistic channel. Shannon's formulation was so fundamental in the sense, he defined the very notion of quantifying information using few basic measures on probability distributions.

In this section some of the key concepts of information, as put forward by Shannon and some of their very essential properties are investigated. Indepth treatment of these concepts and information Theory in general can be gathered from many of the excellent text books in this subject, most notably, [3],[4],[5],[6],[7] [8] and [9]. Shannon's landmark paper [1] itself is an excellent reference on the subject.

There are several key notions in Information Theory. These are basic in the sense that, the whole edifice of Information Theory is built around this. First of such is the notion of entropy.

1.1. **Entropy.**

1.1.1. *Definition of entropy.* Let X be a random variable taking values from a discrete alphabet $\mathcal{X}$ subject to a probability distribution $P_X(x) = \Pr\{X = x\}$ where $x \in \mathcal{X}$. Then the entropy of a (discrete) random variable X is defined as,

$$\begin{aligned} H(X) = H\left(P_X(x)\right) &\triangleq E_{P_X}\left[\log \frac{1}{P_X(x)}\right] \\ &= \sum_{x\in\mathcal{X}} P_X(x) \log \frac{1}{P_X(x)}. \end{aligned} \tag{1}$$

Here E_P is the statistical expectation[1] with respect to the probability distribution P. A further assumption $0 \log 0 = \lim_{t=0} t \log t = 0$ is used for mathematical completeness of the definition. It may be observed that, the usual representation of entropy $H(X)$ is denoted as a function of random variable, even though it is strictly a function of a distribution $P_X(x)$.

Thus, entropy $H(X)$ is the expectation of a random variable $-\log P_X(x)$ with respect to he probability measure P. Since we are considering a discrete random variable, by virtue of $0 \leq P(x) \leq 1$, the function $H(X)$ will be lower bounded by 0. In other words, the entropy is always non-negative. i.e., $H(X) \geq 0$. In general, the upper bound on entropy can be ∞, unless the disribution takes on a countable set of values. The latter assumption is a reasonably one in practice since most of the discrete distributions we come across indeed have only countable number of distinct letters. The easiest example of a countable distribution we could think of is a binary distribution (a single coin flip) with two letters, of probabilities p and $1-p$. The entropy for such a distribution can be easily computed as $p \log p + (1-p)\log(1-p)$. If the alphabet size of the discrete distribution is $|\mathcal{X}|$, the entropy has an uperbound $\log |\mathcal{X}|$.In the binary case, the upper bound thus is $\log 2 = 1$. This rather simple entropy function for a binary case is shown in Fig.1.

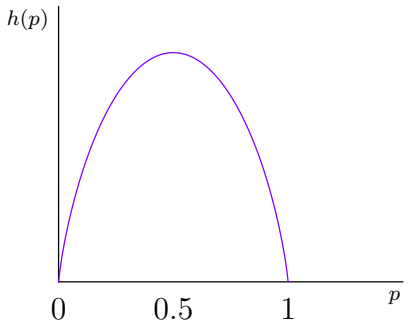

Figure 1. Entropy bounds of a binary distribution: The entropy function $h(p) = p \log p + (1-p)\log(1-p)$ shows general insights into the entropy function of a discrete distribution with a countable alphabet size. If the number of distinct letters that the random variable X take is $\mathcal{X}$, then the maximum value of entropy is $\log |\mathcal{X}|$. The concave nature of the entropy function for a binary distribution shown here also holds true in general, for larger alphabets

[1]Strictly speaking the expectation is E_{P_X} and the the distribution under consideration to be denoted as $P_X(x)$, but partly because of convenience and partly because of the obvious notion, the term X is omitted in the representation.

In most systems that deals with Information Theory, at least two entities are relevant. In a communication system, these are the transmitter (sender) and receiver. We are hence required to consider a pair of random variables not just a single random variable. The two random variables (corresponding to the two entities) are correlated to each other (in the special case they can be independent too). In such a scenario, it is possible to define the joint entropy $H(X,Y)$ between two random variables X and Y. The concept could be extended to an arbitrary number n of random variables $(X_1, X_2, \ldots, X_n)$ with joint entropy $H(X_1, X_2, \ldots, X_n)$.

For two random variables, we can also define the entropy conditioned on an event. In the same vein, we define the averaged (with respect to the distribution of the conditional event) entropy conditioned on an event, known as conditional entropy. The following illustrate the concepts:

The entropy of random variable X conditioned on an event x is defined as,

$$H(X|Y=y) = \sum_{x\in\mathcal{X}} P_{X|Y}(x|Y=y)\log\frac{1}{P_{X|Y}(x|Y=y)} \tag{2}$$

Re-working the above will lead us to

$$H(X|Y=y) = \sum_{x\in\mathcal{X}} P_{X|Y}(x|Y=y)\log\frac{1}{P_{X|Y}(x|Y=y)} \tag{3}$$

$$= E_{P_{X|Y}}\left[\log\frac{1}{P_{X|Y}(x|Y=y)}\right] \tag{4}$$

Expectation of this with respect to $P_Y(y)$ gives us what is known as conditional entropy $H(X|Y)$ between random variables X and Y.

$$H(X|Y) = \mathbb{E}\left[H(X|Y=y)\right] \tag{5}$$

$$= \sum_{y\in\mathcal{Y}} P_Y(y)\sum_{x\in\mathcal{X}} P_{X|Y}(x|Y=y)\log\frac{1}{P_{X|Y}(x|Y=y)} \tag{6}$$

$$= \sum_{y\in\mathcal{Y}}\sum_{x\in\mathcal{X}} P_Y(y)P_{X|Y}(x|Y=y)\log\frac{1}{P_{X|Y}(x|Y=y)} \tag{7}$$

$$= \sum_{y\in\mathcal{Y}}\sum_{x\in\mathcal{X}} P_{X,Y}(x,)\log\frac{1}{P_{X|Y}(x|Y=y)} \tag{8}$$

$$= E_{P_{X,Y}}\left[\log\frac{1}{P_{X|Y}(x|Y=y)}\right] \tag{9}$$

1.1.2. *Additivity of entropy.* A simple additive property exists between entropy, joint entropy and conditional entropies. This is known as the chain rule of entropy. For the two random variable case, it reflects as,

$$H(X|Y) = \sum_{y\in\mathcal{Y}}\sum_{x\in\mathcal{X}} P_{X,Y}(x,)\log\frac{1}{P_{X|Y}(x|Y=y)} \tag{10}$$

$$= \sum_{x\in\mathcal{X},y\in\mathcal{Y}} P_{X,Y}(x,y)\log\frac{1}{P_{X|Y}(x|Y=y)} \tag{11}$$

$$= \sum_{x\in\mathcal{X},y\in\mathcal{Y}} P_{X,Y}(x,y)\log\frac{P_Y(y)}{P_{X,Y}(x,y)} \tag{12}$$

$$= \sum_{x\in\mathcal{X},y\in\mathcal{Y}} P_{X,Y}(x,y)\log\frac{1}{P_{X,Y}(x,y)} - \sum_{x\in\mathcal{X}} P_Y(y)\log\frac{1}{P_Y(y)} \tag{13}$$

$$= H(X,Y) - H(Y) \tag{14}$$

It is easily seen that symmetric property holds (change the random variables X to Y) In summary,

$$H(X,Y) = H(X) + H(Y|X) = H(Y) + H(X|Y). \tag{15}$$

The property can be extended to arbitrary number of random variables to get the chain rule in general.

$$H(X_1, X_2, X_3, \ldots, X_n) = H(X) + H(X_2|X_1) + H(X_3|X_1, X_2) + \ldots + H(X_n|X_1, X_2, X_3, \ldots, X_{n-1})$$

1.2. **Mutual information.** Mutual information between two random variables X and Y is defined as the reduction of entropy of one (say X) given the other (Y). It is denoted as $I(X;Y)$ and is formally,

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

$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y) \\ &= H(X) + H(Y) - H(X,Y) \\ &= \sum_{x\in\mathcal{X}} P_X(x)\log\frac{1}{P_X(x)} + \sum_{y\in\mathcal{Y}} P_Y(y)\log\frac{1}{P_Y(y)} - \sum_{x\in\mathcal{X},y\in\mathcal{Y}} P_{X,Y}(x,y)\log\frac{1}{P_{X,Y}(x,y)} \\ &= \sum_{x\in\mathcal{X},y\in\mathcal{Y}} P_{X,Y}(x,y)\log\frac{P_{X,Y}(x,y)}{P_X(x)P_Y(y)} \\ &= E_{P_{XY}}\left[\log\frac{P_{X,Y}(x,y)}{P_X(x)P_Y(y)}\right]. \end{aligned}$$

By symmetry, the following is true as well:

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

1.3. **Conditional mutual information.**

$$\begin{aligned} I(X;Z|Y) &= \sum_{y\in\mathcal{Y}} P_Y(y) I(X;Z|Y=y) \\ &= \sum_{y\in\mathcal{Y}} P_Y(y) \sum_{x,z} P_{X,Z|Y}(x,z|y)\log\frac{P_{X,Z|Y}(x,z|y)}{P_{X|Y}(x|y)P_{Z|Y}(z|y)}. \end{aligned}$$

1.4. **Inequalities concerning mutual information.**

1.4.1. *simple 3 rv Markov chain.*

$$I(X;Z|Y) \geq 0 \tag{16}$$

and

$$H(X|Y,Z) \leq H(X|Y) \tag{17}$$

and
equality only if $X \to Y \to Z$.

1.4.2. *Markov chain.* For a simple Markov chain

$$X_1 \to X_2 \to X_3 \to \ldots \to X_n, \tag{18}$$

$$I\left(X_1, X_2, X_3, \ldots X_{i-1}; X_{i+1}|X_i\right) = 0. \tag{19}$$

1.4.3. *Independence.* If each component of the random vector

$$\mathbf{X}^n = (X_1, X_2, X_3, \ldots, X_n) \tag{20}$$

is independent from all others,then

$$I(\mathbf{X}^n; \mathbf{Y}^n \geq \sum_{i=1}^{n} I(X_i; Y_i). \tag{21}$$

1.4.4. *Memoryless.* For a memoryless channel, we have

$$I\left(\mathbf{X}^n; \mathbf{Y}^n\right) = \sum_{i=1}^{n} I(X_i; Y_i). \tag{22}$$

Some of these native properties of the basic measures discussed above can be summarized pictorially in Fig. 2.

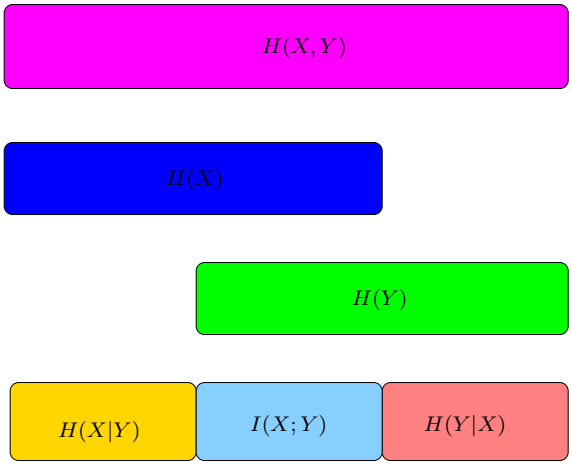

Figure 2. Basic information measures-relationship

2. Inequalities in Information Theory

Information Theory provides fundamental limits on (digital) data transmission and storage. Most of the achievable limits are thus stated in the form of inequalities involving fundamental measures of information such as entropy and mutual information. Such inequalities form a major tool chain to prove many results in information Theory. In a sense, these inequalities separates the possibilities from impossibilities in Information Theory. The study of information expressions and inequalities thus are of paramount importance in solving key results in information Theory.

What constitutes an Information Theoretic inequality? The simple answer to this would be

any expression, linear or non linear involving information measures, on (multiple) random variables.

The information measures are the usual entropy (single, joint, or conditional) and mutual information (including conditional and those involving multiple random variables). Even though it is not impossible to find a non linear expression involving these measures, they are not much of interest in Information Theory. What brings more interest thus are the linear expressions involving the fundamental measures of information. The fundamental informations are also known as Shannon's information measures. We could formally define an information expression f as a linear combination of Shannon's information measures involving a finite number of random variables. For instance, each of the following are valid information expressions:

$$H(X) + 1.2H(Y|Z) + 0.882I(A;B|C)$$
$$I(X;Y) - 3H(X,Y|Z) + H(A|B,C,D) - 2I(L;M|N,Q)$$
$$I(X;Y|Z) - H(Z) - 3H(X,Y).$$

2.1. **Information inequality.** What makes an information inequality then? Any information expression f such that $f \geq 0$ or $f \leq 0$ candidate itself to be called as an information inequality. By definition two information expressions f and g such that $f \geq g$ or $f \leq g$ also make a valid information inequality. Equality is not required to be explicitly stated since it is equivalent to state the condition of both $\geq$ and $\leq$ being true. For example, if $f \geq g$ and $f \leq g$, then it is as good as saying $f = g$.

2.2. **True information inequality.** When can one say an information inequality is true? Since information expressions are functions of information measures, which itself being (measure) functions of distributions, in order for an information inequality to be (always) true, it must hold the inequality true for all possible (probability) distributions (of random variables). In simplified terms, an information inequality f involving information measures of n random variables, is said to be (always) true if,

- The information inequality is true for any possible sets of distribution involving n random variables (joint probability distribution)

Thus an information inequality satisfied for certain selected distributions, but not for all possible distributions cannot be considered as a true information inequality. However, it is possible to have a constraint on certain random variables and state an information inequality, provided the latter is true for all distributions (under the constraint).

Suppose A is a discrete random variable which takes 3 different values (cardinality of the sample space =3). Then we could write,

$$H(A) \leq \log 3$$

Even though the expression is true for the particular choice of A, the information expression is not quite true in general (When the sample space is expanded to have cardinality more than 3, the entropy could have a higher value than $\log 3$). Now consider,

$$I(X;Y|Z) \geq 0$$

is a true information inequality since this is true for any possible distributions of X, Y and Z. On the other hand

$$I(X;Y|Z) \leq 0$$

is not a true information inequality when no further constraints are assumed. However, if a constraint is imposed in the form that X, Z, Y form a Markov chain $X \rightarrow Z \rightarrow Y$ then $I(X;Y|Z) = 0$. Thus, the expression

$$I(X;Y|Z) \leq 0$$

is a true information inequality with the Markov constraint $X \rightarrow Z \rightarrow Y$.

3. Characterizing information inequalities

Given the importance of information inequalities, it is natural to ask this motivating question. Are there ways, if at all possible, to characterize all information inequalities? Raymond Yeung asked this question and found a rather surprising, simple and amazingly elegant way to characterize, almost all information inequalities. His seminal work [2] brought out an interesting framework to characterize and solve a type of inequalities classified as Shannon Type inequalities. He defines Shannon type inequalities are those, which are (directly or indirectly) implied by the *basic inequalities*, which are inequalities that can be expressed as linear combination of non-negative weighted fundamental measures (Shannon's measures) such as entropy and mutual information. It turns out that, most of the inequalities known till date can be classified as Shannon type. The basic inequalities simply refers to the non-negativity of fundamental measures. Because of the possibility of expressing most of the inequalities (all Shannon type) in terms of positive combinations of basic inequalities, the latter is often referred as the *laws of Information Theory*.

It was long conjectured that[13], there could be laws of Information Theory, outside these simple looking basic inequalities. Such inequalities are now classified as non-Shannon type inequalities. This was indeed validated when Yeung came out with examples of such inequalities [2]. This finding proves that, there exist laws in Information Theory, beyond those laid down by Shannon. While the framework for Shannon type gives a direct way to computationally verify any Shannon type inequality, no such methods are known till date for the non-Shannon type. We will study and discuss Yeung's work on Shannon type inequalities.

The distinct difference between Shannon type and non Shannon type inequalities are further discussed in section 9.2

4. Yeung's framework to solve Shannon type inequalities

Raymond Yeung developed a systematic method to verify all Shannon type inequalities. The outline of Yeung's method is listed below. In subsequent sections, more detailed explanations of the concepts described here are provided.

(1) Let $f \geq 0$ be a given information expression. We need to check whether this indeed is a Shannon type inequality. First we claim that any expression can be written in canonical form $f(\mathbf{h}) = \mathbf{b}^T\mathbf{h}$. By this it mean that, the given expression can be written as a linear combination of entropies and joint entropies, weighed by real scalars. For expression involving n distinct random variables, the canonical representation is essentially of the following form:

$$\begin{aligned} f(\mathbf{h}) = \mathbf{b}^T\mathbf{h} \quad &= \quad \lambda_1 H(X_1) + \ldots + \lambda_n H(X_n) + \lambda_{1,2} H(X_1X_2) + \ldots + \\ &\quad \lambda_{1,2,3} H(X_1, X_2, X_3) + \ldots + \lambda_{1,2,3,\ldots,n} H(X_1, X_2, X_3, \ldots, X_n) \end{aligned}$$

where n is the number of distinct random variables involved in the given expression.

(2) Establish the pyramid Γ_n formed by all elemental inequalities. All elemental inequalities reside in Γ_n

(3) Check whether $\Gamma_n = \mathbf{h} : \mathbf{Gh} \geq \mathbf{0}\} \subset \{\mathbf{h} : \mathbf{b}^T\mathbf{h} \geq \mathbf{0}\}$. This is done using the simplex method of optimization in linear programming (see section **??**): Check whether the minimum for the problem statement below is 0

$$\begin{aligned} &\text{minimize } \mathbf{b}^T\mathbf{h} \\ &\text{s.t. } \mathbf{Gh} \geq \mathbf{0} \end{aligned}$$

If yes the inequality indeed is a Shannon type inequality (by virtue of the following fact). If not, the inequality is either not true or perhaps be a non Shannon type which couldn't be characterized. Further tricks are required to validate such inequalities.

(4) $\Gamma_n^* \subset \Gamma_n$. Here Γ_n^* is the region containing constructible expressions. Any constructable expression has to be an elemental inequality.

Given a set of jointly distributed random variables $X_1, X_2, \ldots, X_n$, we can consider entropies of all random variables $H(X_i)$, entropies of all pairs $H(X_i, X_j)$, etc. ($2^n - 1$ entropy values for all nonempty subsets of $\{X_1, X_2, ..., X_n\}$). For every n-tuple of random variables we get a point in $\mathbb{R}^{2^n-1}$, representing entropies of the given distribution. Following [2] we call a point in $\mathbb{R}^{2^n-1}$ constructible if it represents entropy values of some collection of n random variables. The set of all constructible points is denoted by Γ_n^*. The [15], set of entropy values in Γ_n^* is named as *entropic* set.

It is tempting to ask why we require Γ_n at all, when we have the pyramid of constructible points! The simple reason is that it is hard to characterize Γ_n^* for an arbitrary n (for $n \geq 3$, it is not even closed [**?**]). This is where Yeung pulled out his magicians hat to describe a region Γ_n, which can be characterized from basic inequalities. A more feasible (but also highly non-trivial) problem thus, is to describe the closure $\bar{\Gamma}_n^*$ of n of the set Γ_n^*. The set $\bar{\Gamma}_n^*$ n is a convex cone [**?**], and to characterize it we should describe the class of all linear inequalities of the form

$$\begin{aligned} f(h) = b^T h &= \lambda_1 H(X_1) + \ldots + \lambda_n H(X_n) + \lambda_{1,2} H(X_1 X_2) + \ldots + \\ &\quad \lambda_{1,2,3} H(X_1, X_2, X_3) + \ldots + \lambda_{1,2,3,\ldots,n} H(X_1, X_2, X_3, \ldots, X_n) \end{aligned}$$

which are true for any random variables $X_1, X_2, \ldots, X_n$.(λ_i are real coefficients).

One of the other beautiful finding of Yeung's work is bringing in the relationship between the entropy space and a measure space. He brings in a new idea of a one to one correspondence between information measure (what he refer as I-measure) and a signed measure in a measure field. A brief illustration of this is presented in section 5. He uses this mapping to prove some key results in establishing the minimality of representing information expressions in canonical form. The details of its implication are not addressed in thid report, but the concept is illustrated in the next section. In that sense, sections [**?**] and [**?**] are somwhat detached from the genral flow of this document. Interested readers are encouraged to refer [15] for full justification of this useful idea.

5. Measure Theory basics

Yeung establishes a general, one to one correspondence between Set Theory and Shannon's information measures, using which manipulations of random variables can be done, analogous to that of sets. Effectively, one could use properties of set operations and use them to establish equivalent properties of random variables. A rather short description of the concept used in that endeavour is furnished here. Detailed treatment of this can be seen in [15].

The field $\mathbb{F}_n$ generated by sets $s_1, s_2, \ldots, s_n$ is formed by performing sequence of set operations on these sets. The set operations are

(1) complement
(2) union
(3) intersection
(4) difference

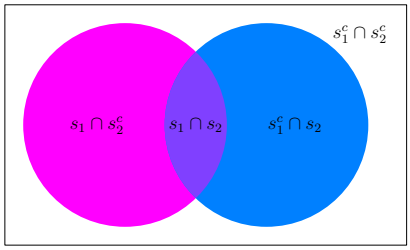

Figure 3. Venn Diagram for two sets s_1 and s_2

As an example, the sets s_1 and s_2 produces 16 elements through the set operations.

- s_1, s_1^c, s_2, s_2^c
- $s_1 \cup s_2, s_1 \cup s_2^c, s_1^c \cup s_2, s_1^c \cup s_2^c$
- $s_1 \cap s_2, s_1 \cap s_2^c, s_1^c \cap s_2, s_1^c \cap s_2^c$

- $s_1 - s_2, s_1 - s_2^c, s_1^c - s_2, s_1^c - s_2^c$

These sixteen elements obtained from the set $\{s_1, s_2\}$ is the field $\mathbb{F}_2$ generated by $\{s_1, s_2\}$. It can be quickly inspected that, not all of them are unique (some can be represented equal or equivalent to other member sets). The number of unique elements of the field are called the *atoms* of the field. They are essentially the sets of the form $\cap_{i=1}^{n} \alpha_i$ where $\alpha_i \in \{s_i, s_i^c\}$

Example: The sets s_1 and s_2 generate $\mathbb{F}_2$, whose atoms are
$\{s_1 \cap s_2, s_1 \cap s_2^c, s_1^c \cap s_2, s_1^c \cap s_2^c\}$
Indeed, any element in the field can be represented as the unions of the subsets of the atoms. In other words, the atoms are the minimal representation of the field itself. The cardinality of the field $\mathbb{F}_2$ is 16 and the number of atoms of $\mathbb{F}_2$ is 4. In general, the number of elements of the field $\mathbb{F}_n$ is 2^{2^n} and the number of atoms are 2^n

It is very helpful to visualize the concept of atoms using *Venn* diagram. The distinct (disjoint) regions of the Venn diagrams are the atoms. All possible unions of these atoms form the field. The simple case of two sets example is shown in Fig.7.

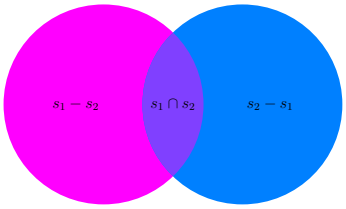

Figure 4. Collapsed field :Venn Diagram for two sets s_1 and s_2

5.1. **Signed measure of a field.** For disjoint $A, B \in \mathbb{F}_n$, a real function μ is called a signed measure if it is set additive, i.e., for disjoint A and $B \in \mathbb{F}_n$,

$$\mu(A \cup B) = \mu(A) + \mu(B) \tag{23}$$

By the definition it implies that

$$\mu(\emptyset) = 0 \tag{24}$$

It can be observed that, a signed measure (again, by definition) μ on $\mathbb{F}_n$ is completely specified by the values on atoms of $\mathbb{F}_n$. Using set additivity, the values of μ on other sets in $\mathbb{F}_n$ can be obtained. For the case of $\mathbb{F}_2$, the 4 values of the signed measures (corresponding to the atoms) are enough to represent all the other 12 values (corresponding to the non atoms in the field).

$$\mu\left(s_1 \cap s_2\right), \mu\left(s_1 \cap s_2^c\right), \mu\left(s_1^c \cap s_2\right), \mu\left(s_1^c \cap s_2^c\right) \tag{25}$$

Values of other elements can be obtained from these measure values. Say for instance $\mu\left(s_1\right)$ can be written as

$$\mu\left(s_1\right) = \mu\left(\left(s_1 \cap s_2\right) \cup \left(s_1 \cap s_2^c\right)\right) \tag{26}$$

$$= \mu\left(s_1 \cap s_2\right) + \mu\left(s_1 \cap s_2^c\right) \tag{27}$$

5.2. **Connection to Shannon's measures.** To establish the connection between Measure Theory (Set Theory to be more precise) and information measures, we have to first associate a set to a random variable.

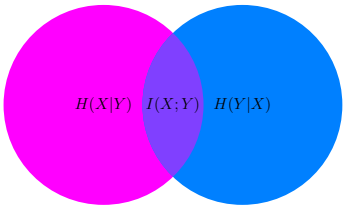

Figure 5. Information diagram for 2 random variables X, Y

Let us consider the simplest case of two random variables X_1 and X_2. We associate two sets, say s_1 and s_2 to the random variables X_1 and X_2 respectively. This set generate a measure field $\mathcal{F}_2$ with cardinality 3 and the atoms. The field

can be expressed conveniently in the form of a Venn diagram. Now let us adopt the following rules to structure the Venn diagram to suit the representation of information measures.

(1) Remove the atom $s_1^c \cap s_2^c$ from the context. Now we are left with 2 atoms. Alternate interpretation of this is: The atom $s_1^c \cap s_2^c$ degenerate to an empty set $\emptyset$. This essentially has the following implication.
(2) Collapse the universe Ω to simply the union of the two sets $s_1 \cup s_2$. Here we force the universal set to shrink into simply the union of non-empty atoms of field $\mathcal{F}_2$ (That is atoms of $\mathcal{F}_2$ excluding $s_1^c \cap s_2^c$). By doing this, we have essentially shrunk the Venn diagram as well (The box region disappeared!)
(3) The new universe now is $s_1 \cup s_2$ and there are 2 non empty atoms which are $\{s_1 \cap s_2, s_1^c \cup s_2, s_1 \cup s_2^c\}$.

The Shannon's information measures for two random variables X_1 and X_2 are,

$$H(X_1), H(X_2), H(X_1|X_2), H(X_2|X_1), H(X_1, X_2), I(X_1; X_2)$$

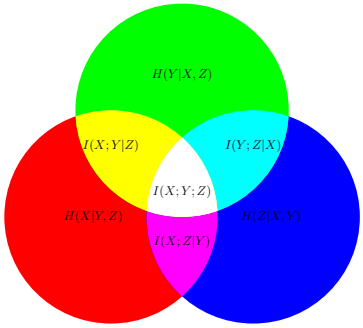

Figure 6. Information diagram for 3 random variables X, Y, Z

Introducing the notation $-$ as in

$$A \cap B^c = A - B$$

we define a signed measure μ by,

$$\begin{aligned} \mu(s_1 - s_2) &= H(X_1|X_2) \\ \mu(s_2 - s_1) &= H(X_2|X_1) \\ \mu(s_2 \cap s_1) &= I(X_1; X_2) \end{aligned}$$

These are the measures on the non-empty[2] atoms of the field $\mathcal{F}_2$. Using the measure property, the measures of other elements of the field can be obtained by addition of these measures on atoms.

For example,

$$\begin{aligned} \mu(s_1 \cup s_2) &= \mu([s_1 - s_2] \cup [s_2 - s_1] \cup [s_1 \cap s_2]) \\ &= \mu(s_1 - s_2) + \mu(s_2 - s_1) + \mu(s_1 \cap s_2) \\ &= H(X_1|X_2) + H(X_2|X_1) + I(X_1; X_2) \\ &= H(X_1, X_2) \\ \mu(s_1) &= \mu([s_1 - s_2] \cup [s_1 \cap s_2]) \\ &= H(X_1|X_2) + I(X_1; X_2) \\ &= H(X_1|X_2) + H(X_1) - H(X_1|X_2) \\ &= H(X_1) \\ \mu(s_2) &= \mu([s_2 - s_1] \cup [s_2 \cap s_1]) \\ &= H(X_2|X_1) + I(X_1; X_1) \\ &= H(X_2|X_1) + H(X_2) - H(X_2|X_1) \\ &= H(X_2) \end{aligned}$$

[2]atoms of $\mathcal{F}_2$, other than $s_1^c \cap s_2^c$

Thus, the measure on all non-empty elements of the field can be summarized as follows:

$$\begin{aligned} \mu\left(s_1 - s_2\right) &= H\left(X_1|X_2\right) \\ \mu\left(s_2 - s_1\right) &= H\left(X_2|X_1\right) \\ \mu\left(s_2 \cap s_1\right) &= I\left(X_1; X_2\right) \\ \mu\left(s_1 \cup s_2\right) &= H\left(X_1, X_2\right) \\ \mu\left(s_1\right) &= H\left(X_1\right) \\ \mu\left(s_2\right) &= H\left(X_2\right) \end{aligned}$$

from this, we could establish the following mapping:

$$\begin{aligned} \mu &\rightarrow H/I \\ \cup &\rightarrow , \\ \cap &\rightarrow ; \\ - &\rightarrow | \end{aligned} \tag{28}$$

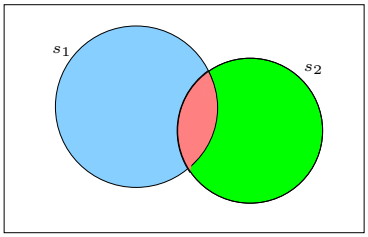

Figure 7. Atoms: Venn diagram of $\mathbb{F}_2$

6. Information measure (I-measure) for arbitrary number of random variables

For a given set of random variables, say n (random variables), the construction of I-measures is merely extending the idea of 2-random variable case.

Let us denote the n random variables as $X_1, X_2, X_3, \ldots, X_n$ and the corresponding to them (respectively) be $s_1, s_2, s_3, \ldots, s_n$.

The universal set Ω is a collapsed version of the conventional universe[3]. In simple terms,

$$\Omega = \bigcup_{i \in \mathcal{N}_n} s_i \tag{29}$$

Where $\mathcal{N}_n$ is,

$$\mathcal{N}_n = \{1, 2, 3, \ldots, n\} \tag{30}$$

Because of the collapsing, the atom formed by the complement intersection $\bigcap_{i \in \mathcal{N}_n} s_i^c$ degenerate to empty set. That is,

$$\bigcap_{i \in \mathcal{N}_n} s_i^c = \left(\bigcap_{i \in \mathcal{N}_n} s_i \right)^c = \Omega^c = \emptyset \tag{31}$$

The cardinality of non empty atoms of $\mathcal{F}_n$ is $2^n - 1$. Extending the idea of two random variable (and two corresponding set scenario) we can claim that, a signed measure μ on $\mathcal{F}_n$ is $2^n - 1$ is fully specified by the measure μ on non empty atoms of $\mathcal{F}_n$. A formal proof of this can be found in [2].

[3]If the universe were not collapsed, the field would also contain the element $\bigcap_{i \in \mathcal{N}_n} s_i^c$. Collapsing the universe can be thought of as the case where in $\bigcap_{i \in \mathcal{N}_n} s_i^c = \emptyset$

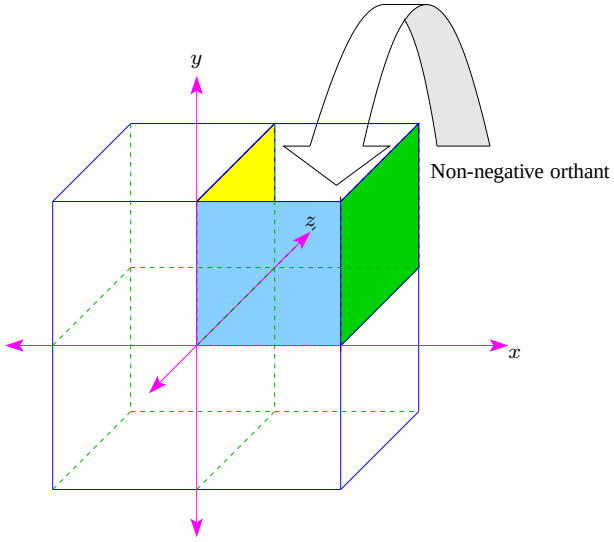

Figure 8. Non negative orthant illustration for 3 dimension

7. Entropy Space

7.1. **Entropy Space $\mathcal{H}_n$: The region Γ^*.** With n random variables, we have $2^n - 1$ joint entropies (including the n entropies of individual random variables).

Examples:

(1) $n = 3$: Let the random variables be X, Y, Z. The non empty joint entropies are

$$H(X), H(Y), H(Z),$$
$$H(X,Y), H(Y,Z), H(X,Z),$$
$$H(X,Y,Z)$$

(2) For $n = 4$,Let the random variables be A, B, C, D, then the non empty joint entropies (15 of them) are

$$H(A), H(B), H(C), H(D),$$
$$H(A,B), H(B,C), H(C,D), H(A,C), H(A,D), H(B,D),$$
$$H(A,B,C), H(B,C,D), H(A,B,D), H(A,C,D),$$
$$H(A,B,C,D)$$

Now, let us consider a set of n random variables. Each of the entropies (and joint entropies) associated with this chosen set of random variables are non negative real values (depending solely on the probability and joint probability distribution of the random variables in hand). If we consider several possible sets of such n random variables, the entropy values could assume many different (some times same as other sets) real values (non negative). Thus for every n random variables we have a $2^n - 1$ tuple of real values.

Now, we think of an Euclidean space of dimension $2^n - 1$. Let the space have co-ordinates labeled as $h_i, i = 1, 2, \ldots, 2^n - 1$. Let us call this space as $\mathcal{H}_n$. The $2^n - 1$ tuple corresponding to a random variable set (of n random variables) is a column vector in $\mathcal{H}_n$. A column vector $h \in \mathcal{H}_n$ is called *entropic* if the $2^n - 1$ tuple represented by h correspond to a valid set of random variables[4]. In other words, when the vector h contains elements (co-ordinate weights) which correspond to joint entropies for any valid random variable set (valid probability distributions) then h is entropic. An example will illustrate this concept:

Example: Let $n = 2$, the entropy space $\mathcal{H}_n$ has co-ordinates h_1, h_2, h_{13}

$$h = \begin{bmatrix} 1 \\ 0.5 \\ 0.25 \end{bmatrix}$$

[4]Yeung in his papers also defined a term entropy function, $H_\Theta(\alpha)$

is not entropic since $H(X) = 1, H(Y) = 0.5$ and $H(X,Y) = 0.25$ does not correspond to a valid entropy measures for any distribution. This can be checked by

$$\begin{aligned} H(X,Y) - H(X) &= H(Y|X) \geq 0 \\ 0.5 - 1 &= H(Y|X) \geq 0 \end{aligned}$$

cant be true. Hence it is not entropic.

The region in the Euclidean space $\mathcal{H}_n$ where h is entropic is of special interest. This region denoted as Γ_n^*. Formally,

$$\Gamma_n^* = \{h \in \mathcal{H}_n : h \quad \text{is entropic}\}$$

Clearly, all entropy measures are non negative, which necessitates that the region Γ_n^* is in the non-negative orthant of the $2^n - 1$ dimensional space $\mathcal{H}_n$. The origin is included in Γ_n^* since all constant n random variables (special case when all the random variables are deterministic[5]) has h an all 0 tuple.

8. Shannon's Information measures in canonical form

All Shannon's information measures (entropies, conditional entropies and mutual informations) can be expressed as a linear combination of entropies and joint entropies. The well known identities to do this translation are

$$\begin{aligned} H(X|Y) &= H(X,Y) - H(Y) \\ H(Y|X) &= H(X,Y) - H(X) \\ I(X;Y) &= H(X) + H(Y) - H(X,Y) \\ I(X;Y|Z) &= H(X,Z) + H(Y,Z) - H(Z) - H(X,Y,Z) \end{aligned}$$

This style of representation in terms of joint (and single) entropies is known as canonical representation of information expressions. Mathematically,

$$f(h) = b^T h \tag{32}$$

Canonical form representation is unique[15].

9. Information Inequalities in elemental form

All information measures formulated by Shannon are non negative measures. These measures, known as Shannon's measures are quantities defined as the entropies,conditional entropies, joint entropies, mutual informations and conditional mutual informations. It is rather rudimentary to check the following basic properties

$$\begin{aligned} H(X) &\geq 0 \\ H(Y) &\geq 0 \\ H(X,Y) &\geq 0 \\ H(X|Y) &\geq 0 \\ I(X;Y) &\geq 0 \\ H(X,Y,Z) &\geq 0 \\ H(X,Y|Z) &\geq 0 \\ I(X;Y|Z) &\geq 0 \end{aligned}$$

These are some of the Shannon's' measures with up to 3 random variables. For any set of random variables, all possible such measures are non-negative. This non negativity of all Shannon's information measures form a set of inequalities known as *basic inequalities*. It may be noted that, the basic inequalities are not unique, in the sense that some of them can be directly inferred from other. This is by virtue of the fact that, Shannon's information measures can itself be written in terms of some or more (linear) combinations of themselves. For instance a Shannon's measure $H(X|Y)$ can also be written as follows:

$$H(X|Y) = H(X|Z,Y) + I(X;Z|Y) \tag{33}$$

Here one information measure is written as sum of two information measures, all of them are Shannon's' information measures.

[5]However contradicting this may be!

9.1. **Elemental Information measures.** An information measures in the form of entropies, conditional entropies, mutual information or conditional mutual information is termed as elemental information measure. More precisely, they are of either of the following form

(1) $H\left(X_i|X_{N_n-i}\right), i \in \mathcal{N}_n$
(2) $I\left(X_i;X_j|X_K\right), i \neq j, K \subset \mathcal{N}_n - \{i,j\}$

where

$$\mathcal{N}_n = \{1,2,3,\ldots,n\}$$

is a set of numbers from 1 to n ($n \geq 2$). X_{N_n-i} refer to string (all of n') of random variables excluding X_i. $X_{N_n-\{i,j\}}$ is a string of random variables, not including X_i and X_j. Note that, $X_i;X_j|X_K$ with $K \subset \mathcal{N}_n - i,j$ refers to any string (including null string) not including X_iX_j. The following example will clarify this.

Example: $H(X_1,X_2)$ can be written as,

$$\begin{aligned} H(X_1,X_2) &= H(X_1) + H(X_2|X_1) \\ &= H(X_1|X_2,X_3) + I(X_1;X_2,X_3) + H(X_2|X_1,X_3) + I(X_2;X_3|X_1) \\ &= H(X_1|X_2,X_3) + I(X_1;X_2) + I(X_1;X_3;X_2) \\ &\quad + H(X_2|X_1,X_3) + I(X_2;X_3|X_1) \end{aligned}$$

In general, for n random variables, total number of elemental measures m of the form $H\left(X_i|X_{\mathcal{N}_n-\{i\}}\right)$ is n and that of the form $I\left(X_i;X_j|X_K\right), i \neq j, K \subset \mathcal{N}_n - i,j$ are

$$\begin{aligned} m &= \binom{n}{2} \times \left[\binom{n-2}{0} + \binom{n-2}{1} + \ldots + \binom{n-2}{n-3} + \binom{n-2}{n-2}\right] \\ &= \binom{n}{2} \times 2^{n-2} \end{aligned}$$

Together, total number of Shannon's information measures in elemental form, for n random variables is

$$m = n + \binom{n}{2} 2^{n-2} \tag{34}$$

Since there are m elemental forms for n random variables, we have m non-negative measures. This is just restating the fact that the elemental forms are always non-negative. This set of m inequalities (≥ 0) compose what is known as *elemental inequalities*. With the example with $n = 3$ we confirm the already known fact that $H(X_1,X_2) \geq 0$ using elemental inequalities.

$$\begin{aligned} H(X_1,X_2) &= \underbrace{H(X_1|X_2,X_3)}_{\geq 0} + \underbrace{I(X_1;X_2)}_{\geq 0} + \underbrace{I(X_1;X_3;X_2)}_{\geq 0} \\ &\quad + \underbrace{H(X_2|X_1,X_3)}_{\geq 0} + \underbrace{I(X_2;X_3|X_1)}_{\geq 0} \\ &\geq 0 \end{aligned}$$

It turns out that, the set of elemental inequalities form a considerable space where in many information inequalities reside. In fact, Yeung uses (and proves) this very own fact to check whether an arbitrary information expression satisfy inequality or not.

9.2. **Elemental inequalities in canonical form.** The $m = n + \binom{n}{2}2^{n-2}$ elemental inequalities can also be expressed in canonical form (with just entropies and joint entropies). This seemingly redundant step is not merely to validate the existence of a canonical form for elemental inequalities. It rather helps us to formulate a good geometrical and subsequently to a linear programming framework. The idea is this: When the elemental inequalities are expressed in canonical form, it become linear inequalities in entropy space $\mathcal{H}_n$. Yeung define a region Γ_n (Note that, Γ_n^* is not quite the same, but there is some relation, which is coming later) within $\mathcal{H}_n$ where these set of inequalities hold.

Consider a simple elemental inequality as an example $I(X_1;X_2)$. The cannonical representation of this would be:

$$\begin{aligned} I(X_1;X_1) &= H(X_1) + H(X_2) - H(X_1X_2) \\ &= \begin{bmatrix} 1 & 1-1 \end{bmatrix} \begin{bmatrix} H(X_1) \\ H_(X_2) \\ H(X_1,X_2) \end{bmatrix} \end{aligned}$$

Similarly, we can express other elemental inequalities involving two random variables in this form. The collection of all such inequalities form a region Γ_2. The concept extended to arbitrary number of random variables n leades to Γ_n. Since this correspond to linear inequalities, they are of the form $\mathbf{Gh} \geq \mathbf{0}$, where $\mathbf{G}$ is a matrix with real elements.

$$\Gamma_n = \{\mathbf{h} : \mathbf{Gh} \geq \mathbf{0}\} \tag{35}$$

So, what does the region Γ_n tell us? Clearly, this is the region which houses all elemental inequalities. We will consider the example with 2 random variables to get the idea right.

Example:Γ_2

There are 3 elemental inequalities ($n = 2, m = n + \binom{n}{2}2^{n-2} = 2 + 1 = 3$) namely, $I(X_1; X_2) \geq 0, H(X_1|X_2) \geq 0$ and $H(X_2|X_1) \geq 0$. The cannonical representation of these three elemental inequalities are,

$$\begin{aligned} I(X_1; X_2) &= H(X_1) + H(X_2) - H(X_1, X_2) \geq 0 \\ H(X_1|X_2) &= -H(X_2) + H(X_1, X_2) \\ H(X_2|X_1) &= -H(X_1) + H(X_2, X_1). \end{aligned}$$

Expressed in matrix representation this states,

$$\begin{bmatrix} I(X_1; X_2) \\ H(X_1|X_2) \\ H(X_2|X_1) \end{bmatrix} = \underbrace{\begin{bmatrix} 1 & 1 & -1 \\ 0 & -1 & 1 \\ -1 & 0 & 1 \end{bmatrix}}_{\triangleq \mathbf{G}} \underbrace{\begin{bmatrix} H(X_1) \\ H(X_2) \\ H(X_1, X_3) \end{bmatrix}}_{\triangleq \mathbf{h}} \geq \mathbf{0}.$$

Thus the region Γ_2 is simply,

$$\Gamma_2 = \{\mathbf{h} : \mathbf{Gh} \geq \mathbf{0}\}. \tag{36}$$

Because of the *linearity* (in linear inequality), it is easy to characterize the region Γ_n, which includes all elemental inequalities (which are equivalent to basic inequalities involving random variables). Since elemental inequalities are satisfied by entropy function of any random variable set (n of them) satisfying $h \in \Gamma_n^*$, it is clear that

$$\Gamma_n^* \subset \Gamma_n.$$

We have established the inclusion relation of Γ_n^* in Γ_n, but we have insufficient clues as to whether they indeed represent two different regions. We are sure Γ_n^* occupy no larger than Γ_n. We are tempted to ask this question here.

Could Γ_n^* and Γ_n be the same?

If they were so, characterizing one implies the other automatically (both ways). In such a case, we could have concluded that all inequalities in Information Theory are derived from the basic inequalities (through elemental inequalities representation) and a formal way to characterize is available through Γ_n. Most of the inequalities found in the earlier stage of Information Theory were of this form. But the story doesnt end there.

It turned out that, there are inequalities which cannot be derived simply from the basic inequalities. That is, the fundamental Shannon measure non-negativity properties alone, do not lead to all inequalities. First such findings were presented by Yeung and Zhang [18], when they discovered an inequality with four random variables. This strongly asserted the conjecture[6] that, indeed there exist inequalities which cannot be characterized simply by Γ_n. Characterizing Γ_n^* is required instead. In other words, there are laws of Information Theory beyond what is ruled by the fundamental Shannon measure non negativity.

The existence of inequalities beyond what originated from basic Shannon measures, necessetiated clasiffication of information inequalities into two types. They are called

(1) **Shannon type inequalities**: These are ineuqalities which are derived from the basic inequalities. Recall that, basic inequalities are nothing but, the non negativity property of Shannon information measures. Inequalities of this class are completely characterized through Γ_n itself.
(2) **Non Shannon type inequalities**: These are inequalities, which cannot be derived just, from the basic inequality postulates. They are governed by further constraints, which are not yet identified. Some inequalities of this type are known to the Information Theory world. To characterize them, Γ_n is inadequate. It is still and open question, on whether there exist a way to characterize Γ_n^*, which would have solved the riddle.

[6]This question was posed by Pippenger [13] as whether there really exist laws beyond the basic ineqaulities?.

We will focus exclusivley on Shannon type inequalities and study on their characterization a little more detail. For a discussion on non-Shannon type inequalities, readers are referred to [2] and [15]. More recent findings on new class of non-Shannon type inequalities can be seen in [19].

10. Characterizing Shannon type inequalities

We realize that, Shannon type inequalities are those, whcih inherited from the fundamental Shannon measures (basic inequalities). Raymond Yeung's framework enables us to do a characterize them. Yeung's trick hinge on the following rules:

(1) Γ_n is a pyramid in the $k = 2^n - 1$ Euclidean space $\mathcal{H}_n$
(2) $\Gamma_n^* \subset \Gamma_n$

All possible measures of random variables (n random variables) are in the region Γ_n^*. Hence, to check the validity of and information expression $f()$ it is enough to check whether the region (pyramid) $\Gamma_n \subset \{\mathbf{h} : f(\mathbf{h}) \geq \mathbf{0}\}$.

If this condition is established, it is automatic that the expression is true in general for all random variables, since

$$\Gamma_n^* \subset \Gamma_n.$$

In essence, the key to check whether an information expression[7] is to check the following

(1) For once, consider the information expression as an algebraic expression in a Euclidean space (of same dimension) and partition the Euclidean space into two. The region where the inequality holds is the region of interest.
(2) Check whether the region (pyramid) Γ_n of all possible information inequalities (elemental inequalities) reside in the region of interest (where the algebraic inequality stays true). If so, we are sure to say that the expression is true for any random variable set. This is because, all possible expressions involving information measures form a region Γ_n^* which is a subset of Γ_n.

So, in principle we know how to characterize Shannon type inequalities. By virtue of the linearity, further insight can be achieved into Γ_n, which will enpower us to see a geometrical view and subsequent formulation as a computational form. The next section discusses the geometry of Γ_n.

11. Geometry of unconstrained information inequalities

It is rather appealing to put a geometric perspective of the information inequality in an entropy space $\mathcal{H}_n$. Remember,$\mathcal{H}_n$ is $\mathbb{R}^{2^n-1}$ space spanned by joint entropies $H(X_1), H(X_2), \ldots, H(X_1, X_2, \ldots, X_n$. We will illustrate this geometrical idea using an example [2].

Let us examine a Shannon type inequality

$$f = I(X_1; X_2) \geq 0$$

First we write this into canonical form as follows:

$$I(X_1; X_2) = \underbrace{H(X_1) + H(X_2) - H(X_1, X_2)}_{\mathbf{b}^T\mathbf{h}} \geq 0$$

where $\mathbf{h} = \begin{bmatrix} H(X_1) & H(X_2) & H(X_1, X_2) \end{bmatrix}^T$ and $\mathbf{b} = \begin{bmatrix} 1 & 1 & -1 \end{bmatrix}^T$

Now we could see that, $\mathbf{b}^T\mathbf{h} \geq 0$ will split the entropy space $\mathcal{H}_n$ into two regions. But this splitting is more of an algebraic splitting without, any assumption on the validity of the tuple $H(X_1), H(X_2), H(X_1, X_2)$, being entropy values of some distribution. In other words, not all points in the half space $\mathbf{b}^T\mathbf{h} \geq 0$ are entropic. On the other hand, not all tuples which are entropic stay within the half space of interest either. We are exposed to two scenarios here:

(1) The region of all tuples $H(X_1), H(X_2), H(X_1, X_2)$ which are entropic is completely inside the half space $\mathbf{b}^T\mathbf{h} \geq 0$. The pyramid which contain all entropic tuple is denoted by Γ_2^*. So, in this case, $\Gamma_2^* \subset \mathbf{b}^T\mathbf{h} \geq 0$. This scenario would qualify to say that, the given inequality is true (for all possible valid distributions). This is pictorially shown in Fig.9
(2) If there exist at least one entropic tuple, which stay outside the half space $\mathbf{b}^T\mathbf{h} \geq \mathbf{0}$, then we are no longer able to say that the expression is true for all valid distributions. In this case, we could say, the expression is not true. Remember, when we say an expression is true, it means the truthfulness for any probability distribution (even one distribution failing disqualifies the expression being called true). This scenario is illustrated in Fig.10

[7]Let us remind ourselves that, information expressions involves Shannon's measures, associated with random variables through their probability distributions

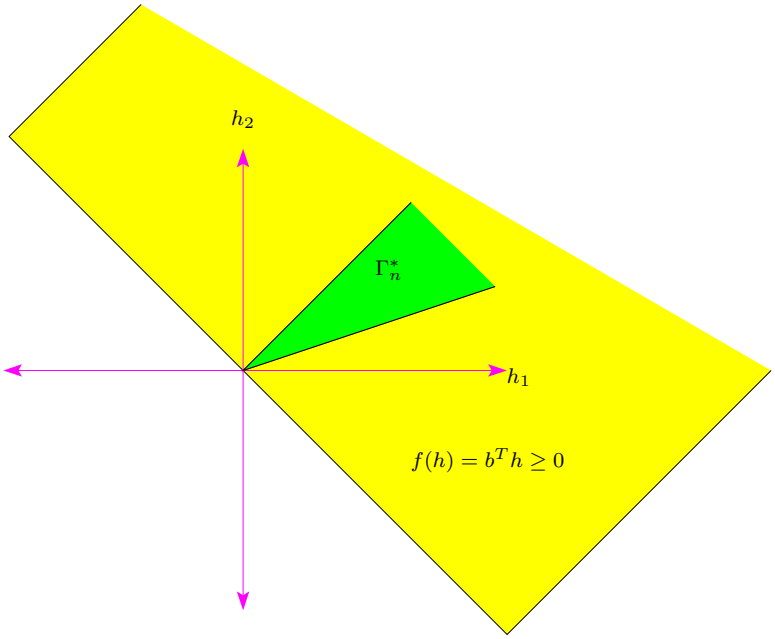

Figure 9. Geometry of unconstrained inequality: Information inequality $f \geq 0$ holds always

.

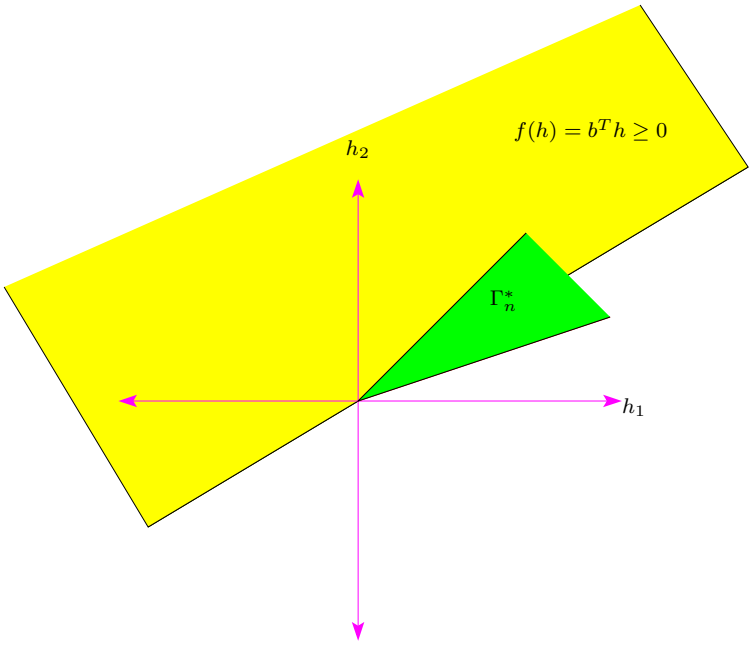

Figure 10. Geometry of unconstrained inequality: Information inequality $f \geq 0$ not necessarily hold always. In this case, it is possible to find a tuple h which is entropic, but reside outside the half space $\mathbf{b}^T\mathbf{h}$

We could extend the example we considered for two random variables to an expression with arbitrary, say n, random variables case. Let us consider a more general information inequality $f \geq 0$. We can write this in canonical form as

$$\begin{aligned} f(h) = \mathbf{b}^T\mathbf{h} &= \lambda_1 H(X_1) + \ldots + \lambda_n H(X_n) + \lambda_{1,2} H(X_1 X_2) + \ldots + \\ &\quad \lambda_{1,2,3} H(X_1, X_2, X_3) + \ldots + \lambda_{1,2,3,\ldots,n} H(X_1, X_2, X_3, \ldots, X_n) \end{aligned}$$

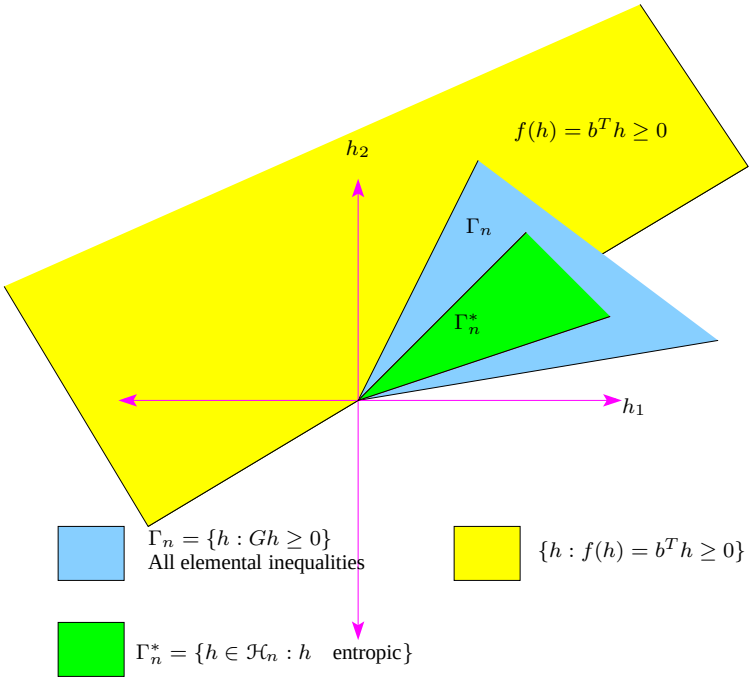

Figure 11. Geometry of unconstrained inequality: Information inequality $f \geq 0$ not necessarily hold always. This is a case where the inequality is not true.

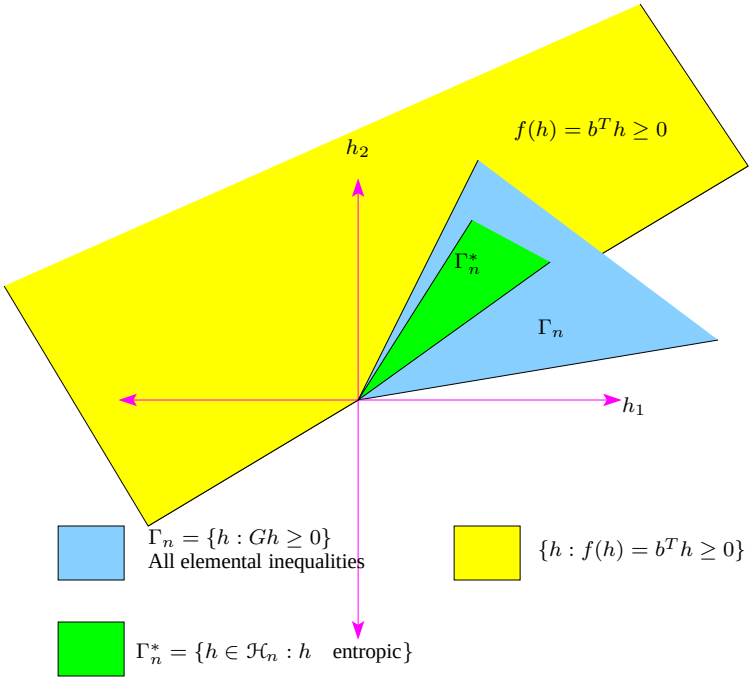

Figure 12. Geometry of unconstrained inequality: Information inequality $f \geq 0$ not necessarily hold always. This is a case of Non Shannon type inequality. Here the inequality is true (since Green region is inside yellow) but not quite a elemental inequality (Blue region partially stay outside yellow region. Better framework needed here to characterize such inequalities.

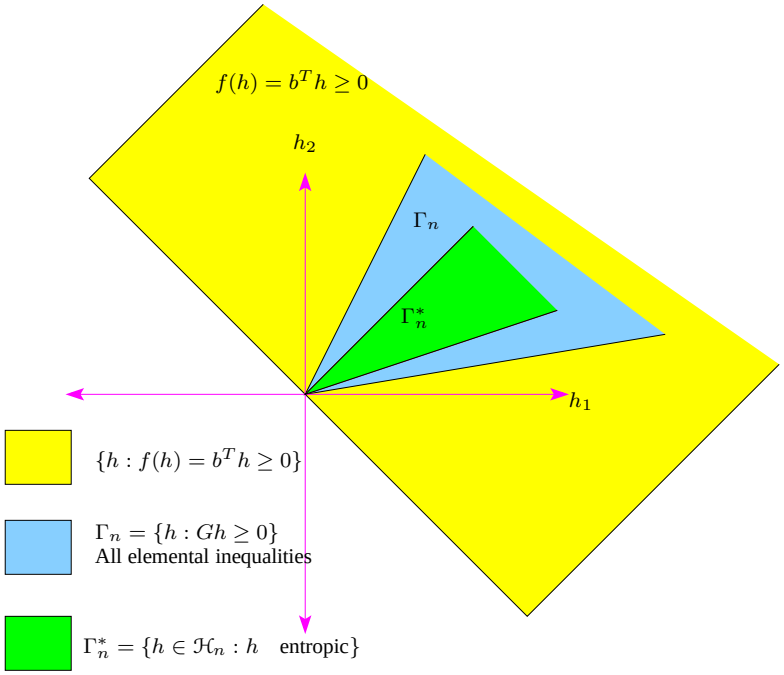

Figure 13. Geometry of unconstrained inequality: Information inequality $f \geq 0$ not necessarily hold always. Here constructible points are completely residing inside the region of Γ_n. Such inequalities can be fully characterized by Γ_n and these are Shannon type inequalities.

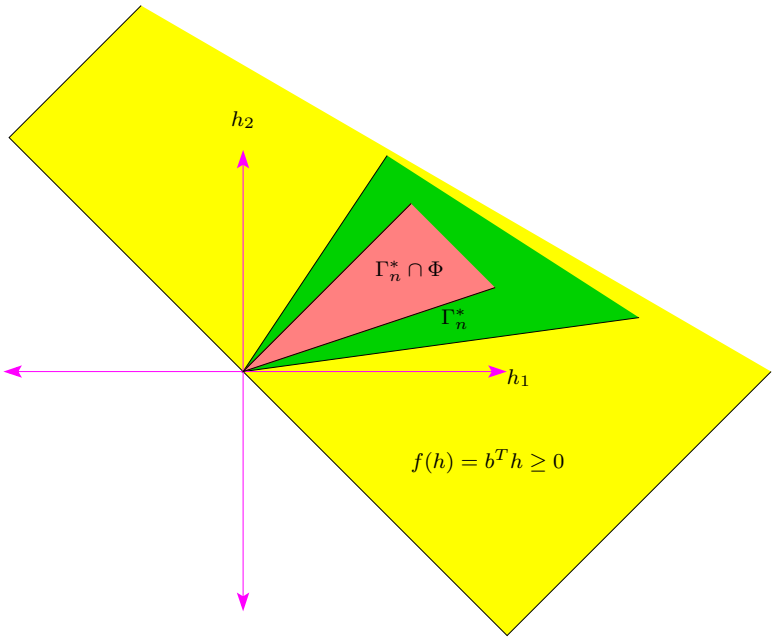

Figure 14. Geometry of constrained inequality: Information inequality $f \geq 0$ holds always. This is the case of constrained inequalities. These are Shannon type inequalities, given the constraints.

We say that, the expression is true (for all distributions) if entropic space stay completely inside the half space determined by the inequality. Formally,

$$f \geq 0 \quad \text{is true iff}^8 \quad \Gamma_n^* \subset \{\mathbf{h} \in \mathcal{H}_n : f(\mathbf{h}) \geq \mathbf{0}\}$$

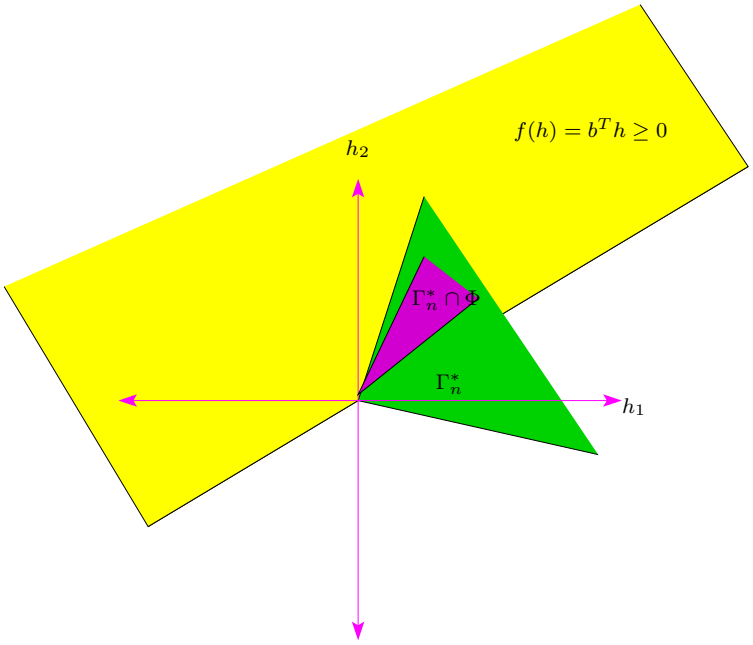

Figure 15. Geometry of constrained inequality: Information inequality $f \geq 0$ holds always, but without constraint, the inequality may not hold always

In principle, this gives a truly complete characterization of unconstrained information inequalities. Unfortunately, it is not that easy to characterize the region Γ_n^*. If we were to do, this, we may have to search for (and construct) the infinite number of possible distributions, which is rather not a viable alternative. However, Yeung had found a way to characterize a larger region named Γ_n which envelope the region Γ_n^*. Here Γ_n refers to the region where all elemental inequalities (Shannon type inequalities) reside. The less tasty part of this sweet method is that, we are no longer able to characterize all information inequalities, but only Shannon type. While majority of the information inequalities are of Shannon type, there exist non Shannon type inequalities as well, as discussed in section 9.2.

Because of the simplicity of the framework, it is indeed possible to formulate the problem into a computational form. This would help us to verify any non Shannon type inequality. Yeung [2] proposed a linear programming framework which could lead to efficient validation of all Shannon type inequalities. We will discuss this next. Detailed discussion on this can be found in [2].

12. Computational method to verify inequalities

Using the framework discussed earlier, it is indeed possible to computationally verify whether any information expression is of Shannon type. The idea, Yeung proposed is briefly discussed here. Only a gist of the idea discussed in [2] is presented here.

12.1. **Linear programming method.** We have seen that, in order to verify whether an information expression $f(h) = b^T h \geq 0$ is Shannon type inequality, we only need to ask the following question:

(1) Is $\Gamma_n \subset \{h : f(h) = b^T h \geq 0\}$?

If the answer is affirmative, then we have the conviction that the expression is indeed a Shannon type inequality. Else, nothing conclusive could be derived at this stage.

A computational procedure to check this condition exist using the well known Linear programming (See section for an elementary treatment on this topic. Readers are referred to the references [**?**] [**?**][**?**][**?**] for more detailed study of this topic.).

For the unconstrained inequality, the problem formulated by Yeung is summarized as follows: Theorem (Yeung): $f(h) = b^T h \geq 0$ is a Shannon type inequality iff the minimum of the problem

$$\begin{aligned} &\text{minimize } \mathbf{b}^T\mathbf{h} \\ &\text{s.t. } \mathbf{Gh} \geq \mathbf{0} \end{aligned}$$

is 0. In this case, the minimum occurs at the origin

13. Constrained inequalities

So far, we have focused on information expressions and inequalities without further constraints. When there is constraints on the joint distributions (of random variables), the dynamics of the information inequalities changes. Information inequalities with such constraints are known as constrained (information) inequalities. The constraints on joint distributions can itself be expressed as linear constraints on the entropies[9]. Following examples illustrate this concept:

(1) X, Y and Z are independent iff

$$H(X,Y,Z) = H(X) + H(Y) + H(Z) \tag{37}$$

$$\begin{aligned} H(X,Y,Z) &= \mathbb{E}\left[\log_2\left(\frac{1}{p_{X,Y,Z}(x,y,z)}\right)\right] \\ &= \mathbb{E}\left[\log_2\left(\frac{1}{p_X(x)p_{Y|X,Z}(y|x,z)p_{Z|X,Y)(z|x,y)}}\right)\right] \\ &= \mathbb{E}\left[\log_2\left(\frac{1}{p_X(x)p_Y(y)p_{Z)(z)}}\right)\right] \\ &= \mathbb{E}\left[\log_2\left(\frac{1}{p_X(x)}\right)\right] + \mathbb{E}\left[\log_2\left(\frac{1}{p_Y(y)}\right)\right] + \mathbb{E}\left[\log_2\left(\frac{1}{p_Z(z)}\right)\right] \\ &= H(X) + H(Y) + H(Z) \end{aligned}$$

(2) Pairwise independence can be expressed through the mutual information. If X, Y, Z are pairwise independent,

$$\begin{aligned} I(X;Y) &= H(X) - H(Y|X) \\ &= H(X) - H(X) \\ &= 0 \\ I(Y;Z) &= H(Y) - H(Z|Y) \\ &= H(Y) - H(Y) \\ &= 0 \\ I(X;Z) &= H(X) - H(Z|X) \\ &= H(X) - H(X) \\ &= 0 \end{aligned}$$

Pairwise equivalence thus necessitates

$$I(X;Y) = I(Y;Z) = I(X;Z) = 0 \tag{38}$$

(3) If $Y = g(X)$ where $g(.)$ is a deterministic function, then $H(X|Y) = 0$. The converse is true as well

(4) Markov Chain $W \to X \to Y \to Z$ implies

$$I(W;Y|X) = 0 \tag{39}$$

$$I(W,X;Z|Y) = 0 \tag{40}$$

(41)

13.1. **Geometrical framework of constrained information inequalities.** Let there be q constraints on distributions, which translates equivalently to q linear constraints on entropies. We could write these equivalent constraints on entropies as a set of q linear equations in the entropy space $\mathcal{F}_n$. But among the q linear equations not all of them may be linearly independent, which means a certain number $r \leq q$ linearly independent equations fully describe the constraints.

$$\mathbf{Qh} = \mathbf{0} \tag{42}$$

where $\mathbf{Q}$ is $q \times k$ matrix ($k = 2^n - 1$).

[9]Here entropies refers to all information measures like entropies, conditional entropies, joint entropies, mutual information, conditional mutual information etc. Also remember that all these information measures can itself be represented in terms of entropies and conditional entropies!

Now the information inequality space shrinks further[10] from the unconstrained space Γ_n^*. Put in other words, the constraints confines the space of information inequality of interest to a linear subspace smaller than

Let

$$\Phi = \{h \in \mathcal{H}_n : Qh = 0\} \tag{43}$$

Now, with this constraint Φ, the expression $f(h) \geq 0$ always holds iff the region $(\Gamma_n^* \cap \Phi) \subset \{h : f(h) \geq 0\}$

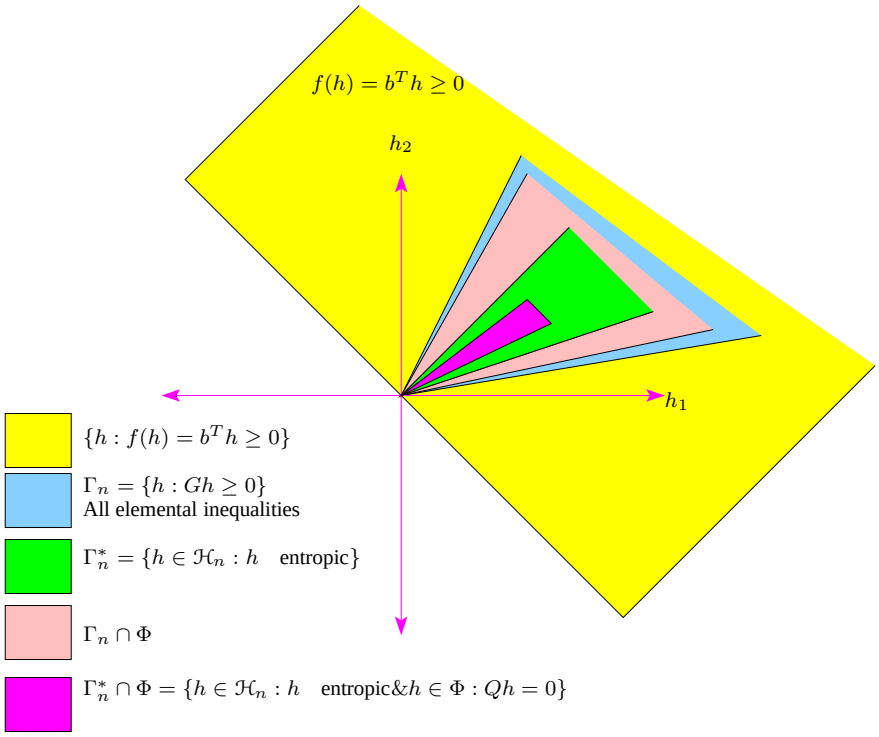

Figure 16. Geometry of constrained inequality: Information inequality $f \geq 0$ holds always. without constraint as well, the inequality hold always

[10]More correctly speaking, the the information inequality space cannot grow beyond Γ_n^*

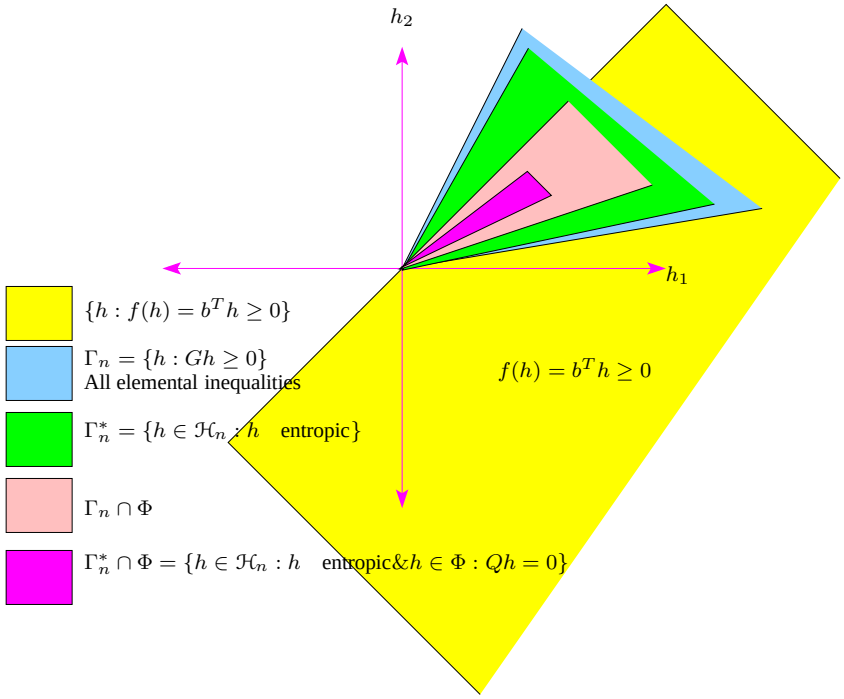

Figure 17. Geometry of constrained inequality: Information inequality $f \geq 0$ holds always. However, without constraint, the inequality is not necessarily true.The region $\Gamma_n \ni f \geq 0, \Gamma_n^* \ni f \geq 0$, but $\Gamma_n^* \cap \Phi \in f \geq 0$. Note that, however this is a non Shannon type inequality since $\Gamma_n \cap \Phi \ni f \geq 0$

14. Linear Programming Basics

Linear programming deals with optimizing a linear cost (objective) function, with linear constraints (inequality constraints as well as equality constraints). Even though it is rather unusual to have a linear cost function, linear programming is often used to solve many problems of practical interest, albeit approximating the cost function to linear.

The number of variables involved in the LP problem can be arbitrary. Since inequality constraints bear a geometrical shape (polyhedron), a more formal definition of LP problem can be stated as follows:

A *linear programming problem*, or *LP*, is a problem of optimizing (maximizing or minimizing) a given linear objective function over some polyhedron. The standard maximization LP, sometimes called the primal problem, is

$$\begin{aligned} &\text{maximize } c^T x \\ \text{(P)} \qquad &\text{s.t. } Ax \leq b \\ &x \geq 0 \end{aligned}$$

Here $c^T x$ is the objective function and the remaining conditions define the polyhedron which is the feasible region over which the objective function is to be optimized. The dual of (P) is the LP

$$\begin{aligned} &\text{minimize } y^T b \\ \text{(D)} \qquad &\text{s.t. } y^T A \geq c^T \\ &y \geq 0 \end{aligned}$$

The linear constraints for a linear programming problems define a convex polyhedron, called the *feasible region* for the problem. The weak duality theorem states that if $\hat{x}$ is feasible (i.e. lies in the feasible region) for (P) and $\hat{y}$ is feasible for (D), then $c^T \hat{x} \leq \hat{y}^T b$. This follows readily from the above:

$$c^T \hat{x} \leq (\hat{y}^T A)\hat{x} = \hat{y}^T (A\hat{x}) \leq y^T b.$$

The strong duality theorem states that if both LPs are feasible, then the two objective functions have the same optimal value. As a consequence, if either LP has unbounded objective function value, the other must be infeasible. It is also possible for both LP to be infeasible.

15. SOFTWARE TOOL TO SOLVE INFORMATION INEQUALITIES

Raymond Yeung and Yan [2] had developed a software package named ITIP [17] to solve all Shannon type inequalities. This software was written in Matlab along with a lexical parser utility yacc. To solve the linear programming problem, they used the LP toolbox of matlab. The tool had its limitations, in terms of license dependability (requires Matlab and Matlab Linear programming toolbox licenses) and computational speed (Matlab is considerably slow compared to a native C program). Besides, the software has become a little outdated in terms of installing (mainly because the dependency packages keep changing). To overcome these, and still to use the seminal work of Yeung, we have developed an all C model software package to solve information inequalities, using the Framework described in [2]. This software is available for free use [**?**]. Essentially three different sets of utilities are available with this package:

(1) A graphical user interface based tool called xiis. One can check any Shannon type inequality with or without constraints by entering the expressions and constraints into the respective entries.
(2) A command line tool named iis, which can take expression and constraints as string arguments.
(3) A file parsing tool which reads a file containing arbitrary number of expressions (one per line) and produces the output in a file.

Some of the enhancements done on the software are listed below:

(1) The entire program, algorithms and computations are written in C language
(2) A parser using lex and yacc to allow different ways to specify random variables. For example, a random variable need not be an English caps letter. Random variable can also be specified as for example,
GamePong, CoinToss_10,X',XX_YY_123
and so on. For instance, it is possible to specify an expression
H(X;X')+2.3 I(John_Lennon_BassLevel;RockFest_1980_Geneva)≥0
where X, X',John_Lennon_BassLevel,RockFest_1980_Geneva are all (valid) random variables.
(3) A graphical user interface tool is built using GtK.
(4) A file based solver is developed using shell script.
(5) To solve linear programming problem we have used the GLPK software tool [**?**], which is available for free under GNU public license.
(6) A speedy version of solving linear programming problem can also be used instead using qsopt [**?**]. We have made softwares using both these versions and they are available for download.

A snapshot view of the tool *xIIS* is shown in Fig.15 and Fig.15

15.1. **Syntax while specifying information expressions and constraints.** In order to use the software, care must be done while specifying the expression and constraints. While the software provide support indicating any wrong syntax, it is worth noting the following notations to be followed, for efficient use of the software. For more detailed specification (with examples) of the software, readers are referred to the xiis user guide [**?**].

(1) Information expression: Information inequality (the one need to be verified) is entered on the top text entry box. Information expressions are linear combinations of any basic measures. The basic information measures can be scaled by real values (can be negative as well). Some examples are:

$$\begin{aligned} I(X;Y) + 2H(A\_1, B') &\geq 0 \\ H(A, B, SnowLevel) - 1.23I(X;Y) - 2H(A|B) &\leq 0 \\ I(X;YY) &= H(X) - H(X|YY) \end{aligned}$$

(2) The information expression must be either and equality or an inequality.
(3) While arbitrary scaling of information measures are allowed, real numbers without associating a measure of random variable is not allowed. For example, it is not allowed to specify $H(X,Y) + 2I(X;Y) + 3 \geq 0$
(4) The constraints are entered in the second entry box. One constraint per line within the entry box is expected
(5) Constraints cannot be inequalities
(6) Constraint could be an equality expression, a Markov chain or independence
(7) Independence is specified by a dot. For instance, to specify three random variables X, Y, Z to be independent, the constraint is specified as
$$X.Y.Z$$
(8) W, X, Y, Z forming a Markov chain $W \to X \to Y \to Z$, the constraint is specified (using a forward slash) as,
$$W/X/Y/Z$$

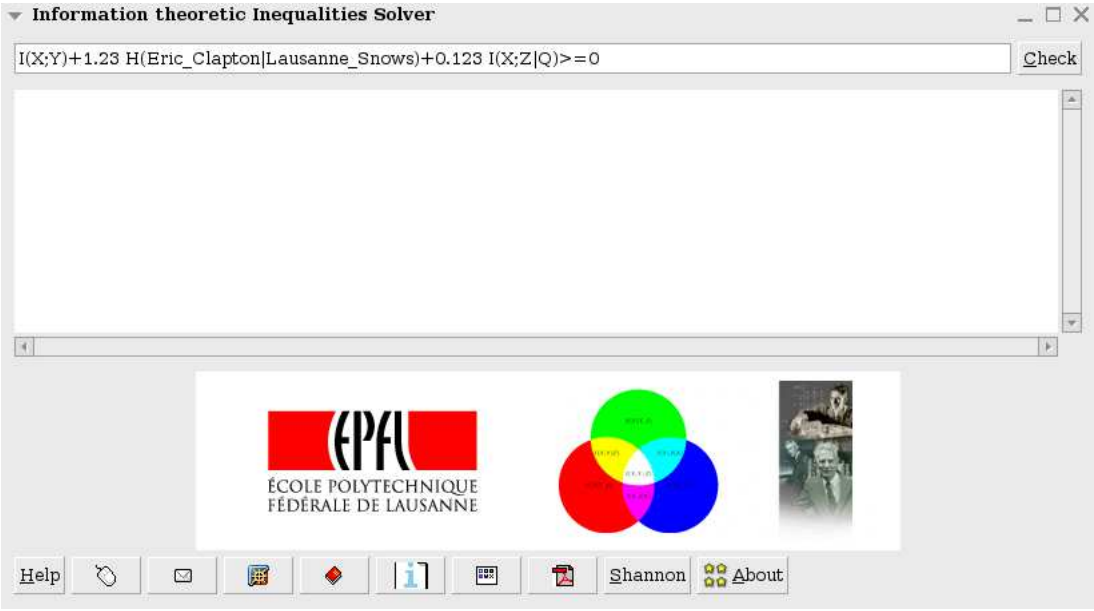

Figure 18. xiis: Information inequality solver main window. The top row entry is where the information expression to be entered. The constraints are to be specified in the text box below. Each constraint must be entered in separate lines. Any number of constraints can be specified. The information expression as well can be arbitrarily long. However the computational time may increase with the number of distinct random variables in the expression and constraints

Figure 19. A brief summary of the xiis software

REFERENCES

[1] C.E.Shannon, "A Mathematical Theory of Communication",Bell System Tech. Journal, Vol. 27, July and October 1948, pp. 379 - 423 and pp. 623 - 656
[2] R. Yeung, A Framework for linear Information Inequalities, IEEE Trans on Information Theory, Vol 43, No.50, Nov 1997
[3] R.G.Gallager, Information Theory and reliable communication, Wiley, New York,1968
[4] T.M.Cover and Joy A. Thomas, Elements of Information Theory, John Wiley and Sons, New York, 1991
[5] T.M.Cover and Joy A. Thomas, Elements of Information Theory, John Wiley and Sons, New York, 2006
[6] R.J. McEliece, The Theory of Information and coding, Addison-Wesley, Reading MA 1977
[7] R.E Blahut, Principles and Practice of Information Theory, Addison-Wesley, Tokyo, 1987
[8] R.E Blahut, Digital Transmission of Information, Addison-Wesley, 1990
[9] R.B.Ash, Information Theory, John Wiley and Sons, Inter-science, New York, 1965
[10] A. Romashchenko, N. Vereshchagin, and A. Shen, Combinatorial Interpretation of Kolmogorov Complexity. Proc. of 15th Annual IEEE Conference on Computational Complexity, July 2000, Florence, Italy, pp. 131-137.
[11] R. W. Yeung and Z. Zhang, A class of non-Shannon type inequalities and their applications. Communications in Information and Systems, 1(2001), pp. 87-100,

[12] T. H. Chan, A combinatorial approach to information inequalities, Comm. Inform. & Syst., 1(2001), pp. 241-253,
[13] N.Pippenger, ”What are the laws of information theory?”, 1986 Special problems on Communication and Computation Conference, Palo Alto, California, Sept, 3-5, 1986
[14] T. H. Chan and R. W. Yeung, On a relation between information inequalities and group Theory, IEEE Trans. Inform. Theory, July 2002.
[15] R. W. Yeung, A First Course in Information Theory, Kluwer Academic/Plenum Publishers, 2002.
[16] J.Matousek and B.Gartner, Understanding and using Linear programming, Springer 2007
[17] R. W. Yeung and Y.O.Yan, Information theoretic Inequality prover (ITIP), http://user-www.ie.cuhk.edu.hk/ĨTIP
[18] R. W. Yeung and Z.Zhang, ”A class of non shannon type information inequalities and their applications”, Comm, Inform & Syst, 1:87-100, 2001
[19] R.Dougherty, C.Freiling, K.Zeger, Six new non Shannon Information Inequalities, IEEE ISIT Seattle, US, July 2006.
[20] G.Strang, Lecture videos: http://math.mit.edu/g̃s/

E-mail address: rethnakaran.pulikkoonattu@epfl.ch
E-mail address: ratnuu@gmail.com