

Research Article

Inequalities Revisited

Raymond Yeung¹

1. Institute of Network Coding, Department of Information Engineering, The Chinese University of Hong Kong, Hong Kong, Hong Kong

In the past over two decades, very fruitful results have been obtained in information theory in the study of the Shannon entropy. This study has led to the discovery of a new class of constraints on the Shannon entropy called non-Shannon-type inequalities. Intimate connections between the Shannon entropy and different branches of mathematics including group theory, combinatorics, Kolmogorov complexity, probability, matrix theory, etc, have been established. All these discoveries were based on a formality introduced for constraints on the Shannon entropy, which suggested the possible existence of constraints that were not previously known. We assert that the same formality can be applied to inequalities beyond information theory. To illustrate the ideas, we revisit through the lens of this formality three fundamental inequalities in mathematics: the AM-GM inequality in algebra, Markov's inequality in probability theory, and the Cauchy-Scharwz inequality for inner product spaces. Applications of this formality have the potential of leading to the discovery of new inequalities and constraints in different branches of mathematics.

Corresponding author: Raymond W. Yeung, whyeung@ie.cuhk.edu.hk

1. Introduction

Inequality defined on the ordered field of real numbers is one of the most fundamental concepts in mathematics. In particular, a *universally quantified* inequality is one that holds for all members in the domain of discourse satisfying certain conditions. Examples are

1. For all $x \in \mathbb{R}$, $e^x \geq 1 + x$.
2. For all $x \in \{y \in \mathbb{R} : y \geq 5\}$, $x \geq 4$.

The latter may simply be written as “If $x \geq 5$, then $x \geq 4$ ”. Essentially, any inequality in mathematics that bears a name is a universally quantified inequality, for example, the AM-GM inequality, the Cauchy-Schwarz inequality, Jensen's inequality, and Minkowski's inequality. The list goes on and on.

The main contribution of this paper is to introduce a framework that gives a geometrical interpretation of such inequalities. This framework has its origin in information theory ^[1], specifically in the study of inequalities on the Shannon entropy (or simply entropy when there is no ambiguity) in the late 1990s ^[2]. This study has led to the discovery of so-called *non-Shannon-type* entropy inequalities, namely inequalities on the Shannon entropy beyond what were known before then (collectively called *Shannon-type* entropy inequalities). Subsequently, intimate relations between the Shannon entropy and different branches of mathematics including group theory, combinatorics, Kolmogorov complexity, probability, matrix theory, etc, have been established. Inspired by this development, there has also been a wave of pursuit of new inequalities on the von Neumann entropy, a generalization of the Shannon entropy to the quantum case.

We assert that this geometrical framework which has led to very fruitful results in the study of entropy inequalities can also be applied to general universally quantified inequalities. In this paper, we will develop the concepts starting with a very simple example, and then apply the concepts to increasingly elaborate examples. The results are presented in a logical order instead of the chronological order that the results were obtained.

The rest of the paper is organized as follows. In Section 2, we discuss the well-known inequality of arithmetic and geometric means (AM-GM inequality) in algebra, and show that the AM-GM inequality completely characterizes the relation between the AM and GM of a finite collection of nonnegative numbers. In Section 3, we discuss Markov's inequality in probability theory. We show that for a nonnegative random variable T and a nonnegative value c , Markov's inequality essentially completely characterizes the relation between the two quantities $E[T]$ (expectation of T) and $\Pr\{T \geq c\}$. In Section 4, we discuss the Cauchy-Schwarz inequality for real inner product spaces. We show that for two vectors $\mathbf{u}, \mathbf{v} \in V$, where V is a real inner product space, the Cauchy-Schwarz inequality completely characterizes the relation among the three quantities $\langle \mathbf{u}, \mathbf{u} \rangle$, $\langle \mathbf{v}, \mathbf{v} \rangle$, and $\langle \mathbf{u}, \mathbf{v} \rangle$ if and only if $\dim(V)$ (the dimension of V) is at least 2. Nevertheless, there exists no inequality on the quantities $\langle \mathbf{u}, \mathbf{u} \rangle$, $\langle \mathbf{v}, \mathbf{v} \rangle$, and $\langle \mathbf{u}, \mathbf{v} \rangle$ that holds for all inner product space V (regardless of the value of $\dim(V)$) which is not implied by the Cauchy-Schwarz inequality. The results in Sections 2 to 4 are new to our knowledge. In Section 5, we give an exposition of the study on entropy inequalities in information theory since the late 1990s. We also briefly discuss the relations between the Shannon entropy and network coding, conditional independence of random variables, finite groups, positive semi-definite matrices, Kolmogorov complexity, and quantum mechanics. The paper is concluded in Section 6.

2. The AM-GM Inequality

The inequality of arithmetic and geometric means, or the AM-GM inequality in brief, is elementary in algebra. In this section, we use this very simple inequality as an example to illustrate the concepts we will develop in this work.

The arithmetic mean and geometric mean of a finite list of nonnegative numbers x_1, x_2, \dots, x_n are

$$\text{AM} = \frac{1}{n}(x_1 + \dots + x_n)$$

and

$$\text{GM} = \sqrt[n]{x_1 \cdot \dots \cdot x_n},$$

respectively. The AM-GM inequality says that

$$\text{AM} \geq \text{GM}. \tag{1}$$

For a collection of proofs of the AM-GM inequality, we refer the reader to [\[3\]](#). Throughout this section, we will use AM and GM to denote the arithmetic mean and the geometric mean of some finite list of nonnegative numbers, respectively.

Traditionally, the AM-GM inequality is interpreted as either a lower bound on the AM or an upper bound on the GM. Here, we take a somewhere different view on the AM-GM inequality that will be elaborated in the rest of the section.

For a finite list of nonnegative numbers, the AM and GM are quantities of interest, and we are interested in the relation between these two quantities. The AM-GM inequality is a characterization of this relation. Since $x_i \geq 0$ for all i , we immediately have $\text{AM} \geq 0$ and $\text{GM} \geq 0$. These inequalities come directly from the setup of the problem. In fact, from $\text{GM} \geq 0$ and $\text{AM} \geq \text{GM}$, we can obtain $\text{AM} \geq 0$. Therefore, $\text{AM} \geq 0$ is redundant.

We now introduce a geometrical framework for understanding the relation between the quantities AM and GM. Let a and g be the coordinates of \mathbb{R}^2 , the 2-dimensional Euclidean space, where a and g correspond to AM and GM, respectively. Define the region

$$\Upsilon = \{(a, g) \in \mathbb{R}^2 : g \geq 0 \text{ and } a \geq g\}.$$

where in the above, $g \geq 0$ and $a \geq g$ correspond to $\text{GM} \geq 0$ and $\text{AM} \geq \text{GM}$, respectively. See Figure 1 for an illustration of Υ .

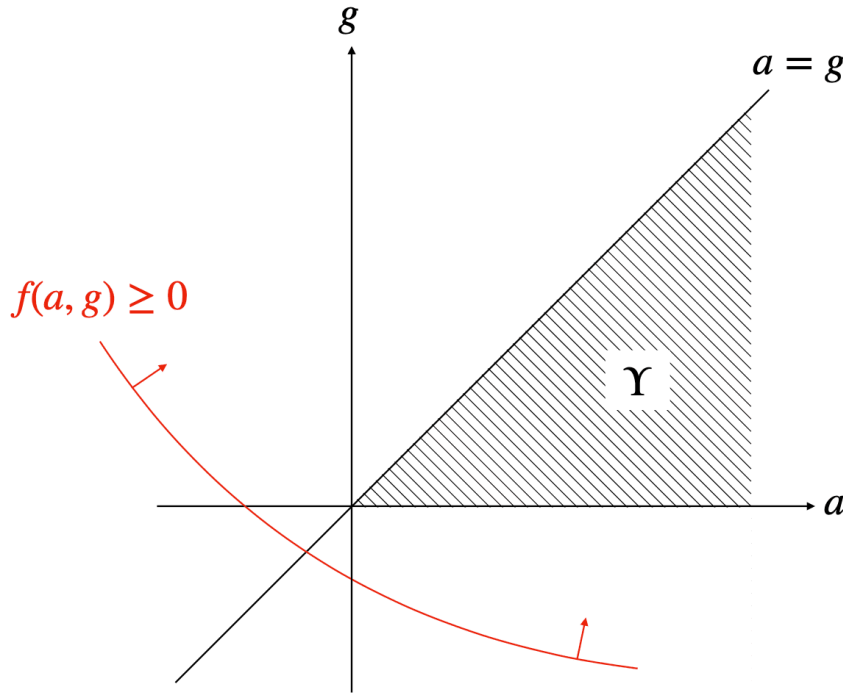


Figure 1. The region Υ in \mathbb{R}^2 and an inequality $f(a, g) \geq 0$ implied by $g \geq 0$ and $a \geq g$.

We now ask a very basic question: Are there constraints on AM and GM other than $GM \geq 0$ and $AM \geq GM$? As we will see, the answer to this question hinges on the next proposition.

Proposition 1. For any $(a, g) \in \Upsilon$, there exist $x, y \geq 0$ such that

$$a = \frac{x + y}{2} \quad \text{and} \quad g = \sqrt{xy},$$

i.e., a and g are the AM and GM of the list of nonnegative numbers x, y , respectively.

Proof. Consider a fixed ordered pair $(a, g) \in \Upsilon$. Since $(a, g) \in \Upsilon$, we have $g \geq 0$ and $a \geq g$. Let

$$a = \frac{x + y}{2} \quad \text{and} \quad g = \sqrt{xy}, \tag{2}$$

where $x, y \in \mathbb{R}$ are unknowns. Here, x and y can be obtained by solving the above simultaneous equations, which is elementary. Note that if $g < 0$, then $g = \sqrt{xy}$ in the above cannot be satisfied, but this is not the case since $(a, g) \in \Upsilon$. The solution to (2) is

$$x = a \pm \sqrt{a^2 - g^2} \quad \text{and} \quad y = a \mp \sqrt{a^2 - g^2},$$

where $a^2 - g^2 \geq 0$ since $a \geq g$. Therefore, x and y are always real. Finally, it follows from $\sqrt{a^2 - g^2} \leq a$ that x and y are always nonnegative. The proposition is proved. \square

An ordered pair (a, g) is an *achievable pair*, or simply *achievable*,¹ if a and g are respectively the AM and GM of some finite list of nonnegative numbers. The next theorem is a consequence of Proposition 1.

Theorem 1. *An ordered pair $(a, g) \in \mathbb{R}^2$ is achievable if and only if $(a, g) \in \Upsilon$.*

Proof. Proposition 1 implies that every ordered pair $(a, g) \in \Upsilon$ is achievable. On the other hand, if an ordered pair (a, g) is achievable, then $(a, g) = (AM, GM)$ where AM and GM are respectively the arithmetic mean and geometric mean of some finite list of nonnegative numbers. Therefore, $g = GM \geq 0$ and $a = AM \geq GM = g$, i.e., $g \geq 0$ and $a \geq g$, implying that $(a, g) \in \Upsilon$. Hence, an ordered pair (a, g) is achievable if and only if $(a, g) \in \Upsilon$. The theorem is proved. \square

Theorem 1 says that Υ is precisely the set of all achievable pairs, thus completely characterizing the relation between the arithmetic and geometric means of a finite list of nonnegative numbers. Since Υ is defined by $g \geq 0$ (corresponding to $GM \geq 0$) and $a \geq g$ (corresponding to the AM-GM inequality), where the former comes directly from the setup of the problem and can be regarded as given, we say that the AM-GM inequality completely characterizes the relation between the AM and GM of a finite list of nonnegative numbers.

An inequality in the quantities AM and GM has the general form

$$f(AM, GM) \geq 0, \tag{3}$$

where $f : \mathbb{R}^2 \rightarrow \mathbb{R}$.² For example, if $f(a, g) = a - g$, then (3) becomes the AM-GM inequality. Let

$$R_f = \{(a, g) \in \mathbb{R}^2 : f(a, g) \geq 0\}$$

be the region in \mathbb{R}^2 induced by $f \geq 0$. If (3) is satisfied for all finite lists of nonnegative numbers, we say that the inequality is *valid*.

Theorem 2. *The inequality (3) is valid if and only if*

$$\Upsilon \subset R_f. \tag{4}$$

Proof. We first prove the “if” part. Assume that (4) holds. Consider any finite list of nonnegative numbers and let AM and GM be the arithmetic mean and geometric mean, respectively. Then (AM, GM) is achievable, and so by Theorem 1, $(AM, GM) \in \Upsilon$. Then by (4), $(AM, GM) \in R_f$, implying that $f(AM, GM) \geq 0$. This shows that the inequality (3) is valid.

Next, we prove the “only if” part by contradiction. Assume that the inequality (3) is valid, i.e., $f(AM, GM) \geq 0$ is satisfied by all finite lists of nonnegative numbers, but

$$\Upsilon \not\subset \{(a, g) \in \mathbb{R}^2 : f(a, g) \geq 0\}.$$

Then there exists an ordered pair $(a_0, g_0) \in \Upsilon$ such that $f(a_0, g_0) < 0$. Since $(a_0, g_0) \in \Upsilon$, by Theorem 1, (a_0, g_0) is achievable, which means that $(a_0, g_0) = (AM^*, GM^*)$, where AM^* and GM^* are respectively the arithmetic mean and geometric mean of some finite list of nonnegative numbers. In other words, we have $f(AM^*, GM^*) = f(a_0, g_0) < 0$, which is a contradiction to the assumption that $f(AM, GM) \geq 0$ is satisfied by all finite lists of nonnegative numbers. The theorem is proved. \square

Theorem 2 gives a complete characterization of all valid inequalities in AM and GM. Specifically, $f(AM, GM) \geq 0$ is a valid inequality in AM and GM if and only if $\Upsilon \subset R_f$, the region induced by $f \geq 0$, is an outer bound on Υ . This is illustrated in Figure 1.

In Theorem 2, the set inclusion in (4) is equivalent to

$$\left. \begin{array}{l} g \geq 0 \\ a \geq g \end{array} \right\} \Rightarrow f(a, g) \geq 0.$$

Upon replacing the dummy variables a and g by AM and GM, respectively, the above becomes

$$\left. \begin{array}{l} GM \geq 0 \\ AM \geq GM \end{array} \right\} \Rightarrow f(AM, GM) \geq 0, \quad (5)$$

meaning that any valid inequality in AM and GM is implied by the inequalities $GM \geq 0$ and $AM \geq GM$. Hence, we conclude that there exists no inequality in AM and GM other than these two inequalities. As discussed, since the inequality $GM \geq 0$ comes directly from the setup of the problem and can be regarded as given, in view of (5), we say that the AM–GM inequality is *sharp*.

We end this section with a remark on the *tightness* of a valid inequality. Let $f(AM, GM) \geq 0$ be a valid inequality. By Theorem 2, the set inclusion in (4) holds. If $f(AM, GM) \geq 0$ is tight, then $f(AM^*, GM^*) = 0$ for some achievable pair (AM^*, GM^*) , which by Theorem 1 is in Υ . If the boundary of the region R_f is equal to

$$\{(a, g) \in \mathbb{R}^2 : f(a, g) = 0\},$$

then (AM^*, GM^*) is in Υ as well as on the boundary of R_f . This implies that the boundary of R_f touches the region Υ at the point where $f \geq 0$ is tight, providing a geometrical interpretation of the tightness of a valid inequality.

As an example, the inequality $2AM \geq GM$ is equivalent to $f(AM, GM) \geq 0$ with $f(a, g) = 2a - g$. It is a valid inequality because $\Upsilon \subset R_f$. Moreover, $2AM \geq GM$ is tight for the list of nonnegative number, 0, with $(AM^*, GM^*) = (0, 0)$. Accordingly, the boundary of R_f , namely the set

$$\{(a, g) \in \mathbb{R}^2 : 2a = g\},$$

touches the region Υ at the origin. This is illustrated in Figure 2.

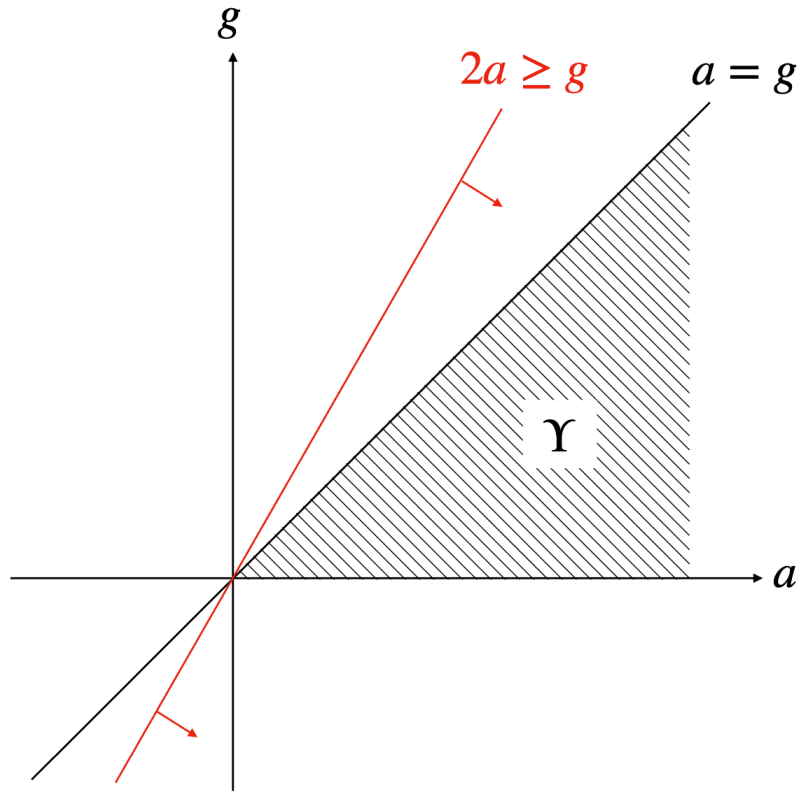


Figure 2. An illustration of the region R_f for $f = 2a - g$.

3. Markov's Inequality

In probability theory, Markov's inequality asserts that for a nonnegative random variable T and any fixed $c > 0$,

$$\Pr\{T \geq c\} \leq \frac{E[T]}{c}, \quad (6)$$

where $\mathbb{E}[T]$ denotes the expectation of T . Here, $\Pr\{T \geq c\}$ and $\mathbb{E}[T]$ are two quantities of interest for any nonnegative random variable T , and we are interested in the relation between them. Markov's inequality gives a characterization of this relation.

Let c in (6) be fixed, and let F_T denote the probability distribution of T , i.e., $F_T(t) = \Pr\{T \leq t\}$. The following is a proof of (6):

$$\begin{aligned} E[T] &= \int_{t \geq 0} t dF_T(t) \\ &= \int_{0 \leq t < c} t dF_T(t) + \int_{t \geq c} t dF_T(t) \\ &\stackrel{i)}{\geq} \int_{t \geq c} t dF_T(t) \\ &\stackrel{ii)}{\geq} c \int_{t \geq c} dF_T(t) \\ &= c \Pr\{T \geq c\}. \end{aligned}$$

Then (6) is obtained by dividing both sides about by c . In the above, the inequality i) is tight if and only if

$$\int_{0 \leq t < c} t dF_T(t) = 0, \tag{7}$$

or

$$\Pr\{0 < T < c\} = 0.$$

Note that although $\Pr\{T = 0\}$ can be strictly positive, it would not in any case make any contribution to (7). On the other hand, the inequality ii) is tight if and only if

$$\Pr\{T > c\} = 0,$$

or equivalently,

$$\Pr\{T \geq c\} = \Pr\{T = c\}. \tag{8}$$

If (6) is tight, i.e., i) and ii) are tight simultaneously, then F_T can only have two point masses, one at 0 and the other at c , with

$$\Pr\{T = 0\} + \Pr\{T = c\} = 1.$$

As a sanity check, for this distribution, from (8), we have

$$E[T] = c \cdot \Pr\{T = c\} = c \cdot \Pr\{T \geq c\},$$

or

$$\Pr\{T \geq c\} = \frac{E[T]}{c}. \quad (9)$$

Thus (6) indeed holds with equality.

Note that the above discussion holds regardless of the value of $\Pr\{T \geq c\}$ (which can be any number between 0 and 1). Thus Markov's inequality can hold with equality for all values of $\Pr\{X \geq c\}$.

We can obtain further insight on Markov's inequality by means of a geometrical framework similar to the one discussed in Section 2 for the AM-GM inequality. Continue to assume that $c > 0$ is fixed. Let p and m be real numbers such that $p = \Pr\{T \geq c\}$ and $m = \mathbb{E}[T]$ for some nonnegative random variable T . Then from (6), we have $m \geq cp$. Since p is a probability, we also have $0 \leq p \leq 1$. Now regard p and m as the two coordinates in \mathbb{R}^2 , and define the region

$$\Psi_c = \{(p, m) \in \mathbb{R}^2 : 0 \leq p \leq 1 \text{ and } m \geq cp\}.$$

See Figure 3.

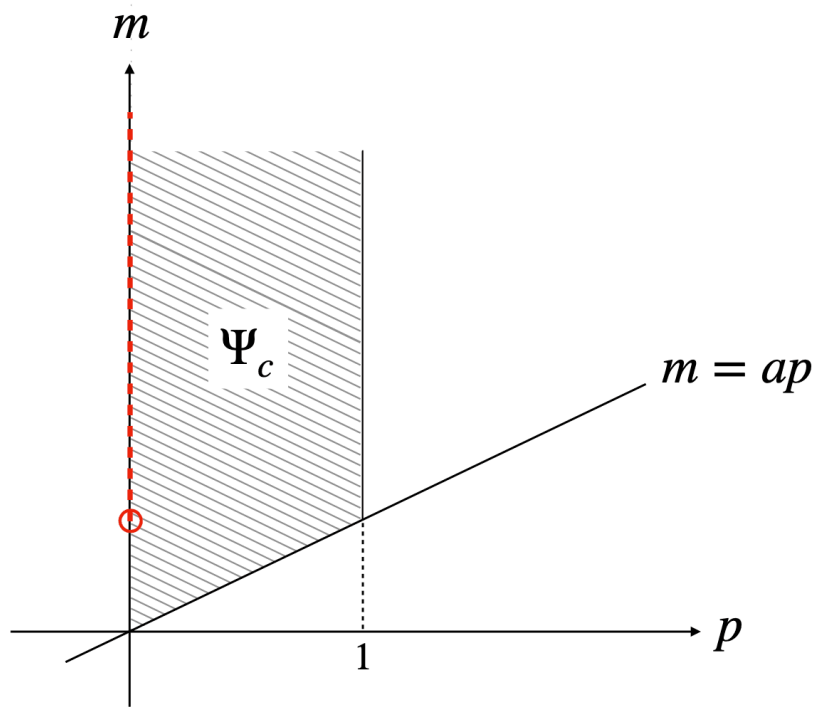


Figure 3. The region Ψ_c in \mathbb{R}^2 .

For real numbers p and m , if there exists a nonnegative random variable T such that $p = \Pr\{T \geq c\}$ and $m = E[T]$, we say that the ordered pair $(p, m) \in \mathbb{R}^2$ is *achievable* (by the random variable T). Note that

an ordered pair in \mathbb{R}^2 may be achievable by more than one nonnegative random variable, i.e.,

$$(p, m) = (\Pr\{T \geq a\}, E[T]) = (\Pr\{T' \geq a\}, E[T'])$$

where T and T' have different probability distributions.

We will show that almost all ordered pairs in Ψ_c are achievable. We first define the *achievable region*

$$\Psi_c^* = \{(p, m) \in \mathbb{R}^2 : (p, m) \text{ is achievable}\}.$$

Since for any nonnegative random variable T , $0 \leq \Pr\{T \geq c\} \leq 1$ and Markov's inequality is satisfied, if (p, m) is achievable, then $(p, m) \in \Psi_c$. Consequently, $\Psi_c^* \subset \Psi_c$.

Theorem 3. $\Psi_c^* = \Psi_c \setminus \{(0, m) : m \geq c\}$.

Proof. First, we prove that $\Psi_c^* \subset \Psi_c \setminus \{(0, m) : m \geq c\}$. Since $\Psi_c^* \subset \Psi_c$, we only need to show that $(0, m)$ is not achievable for all $m \geq c$. When $p = 0$, we have $\Pr\{T \geq c\} = p = 0$, or $\Pr\{0 \leq T < c\} = 1$. This implies that $m = E[T] < c$. In other words, every ordered pair $(0, m)$ where $m \geq c$ is not achievable. Now, we prove that $\Psi_c \setminus \{(0, m) : m \geq c\} \subset \Psi_c^*$. For any $0 < p \leq 1$, consider any $(p, m) \in \Psi_c$ and the following probability mass function for a random variable T :

$$\Pr\{T = 0\} = 1 - p \quad \text{and} \quad \Pr\{T = m/p\} = p.$$

Since $m \geq cp$, we have $m/p \geq c$, and so

$$\Pr\{T \geq c\} = \Pr\{T = m/p\} = p,$$

and $E[T] = p(m/p) = m$. Therefore, (p, m) is achieved by the random variable T as constructed. Note that unless $m = cp$, (p, m) can always be achieved by more than one probability distribution.

It remains to prove that every ordered pair $(0, m)$ with $0 \leq m < c$ is achievable. This can be done by noting that such an ordered pair can be achieved by any random variable T with $\Pr\{T = m\} = 1$. The theorem is proved. \square

The region Ψ_c is defined by Markov's inequality together with the constraint $0 \leq p \leq 1$ which comes from the setup of the problem, and we have shown that for any fixed $c > 0$, every ordered pair in Ψ_c except for a region with Lebesgue measure 0 (namely the region $\{(0, m) : m \geq c\}$) is achievable by some random variable T . Specifically:

- When $\Pr\{T \geq c\} > 0$, Markov's inequality, namely

$$E[T] \geq c \cdot \Pr\{T \geq c\}, \tag{10}$$

which gives a lower bound on $E[T]$, is the only constraint on $E[T]$ in terms of $\Pr\{T \geq c\}$.

- When $\Pr\{T \geq c\} = 0$, Markov's inequality as in (10), which becomes $E[T] \geq 0$, continues to be valid.

However, we also have

$$E[T] < c. \tag{11}$$

Combining (10) and (11), we have

$$0 \leq E[T] < c.$$

See Figure 3 for an illustration. Since every ordered pair in Ψ_c except for a region with Lebesgue measure 0 is achievable, or equivalently, $\overline{\Psi_c^*} = \Psi_c$, we say that Markov's inequality almost completely characterizes the relation between $\Pr\{T \geq c\}$ and $E[T]$.

Now consider an inequality in the quantities $\Pr\{T \geq c\}$ and $E[T]$:

$$f(\Pr\{T \geq c\}, E[T]) \geq 0, \tag{12}$$

where $f : \mathbb{R}^2 \rightarrow \mathbb{R}$. For example, if $f(m, p) = m - cp$, then the inequality in (12) becomes (6), namely Markov's inequality. Let

$$R_f = \{(p, m) \in \mathbb{R}^2 : f(p, m) \geq 0\}$$

be the region in \mathbb{R}^2 induced by $f \geq 0$.

If (12) holds for all nonnegative random variable T , we say that the inequality is *valid*. Markov's inequality is such an example. The fundamental importance of the achievable region Ψ_c^* is explained in the next theorem, which asserts that an inequality $f(\Pr\{T \geq c\}, E[T]) \geq 0$ is valid if and only if $\Psi_c^* \subset R_f$. This implies that the region Ψ_c^* completely characterizes all valid inequalities of the form (12).

Theorem 4. *The inequality (12) is valid if and only if $\Psi_c^* \subset R_f$.*

Proof. We first show that if $\Psi_c^* \subset R_f$, then (12) holds for all nonnegative random variable T . Assume that $\Psi_c^* \subset R_f$. Consider any nonnegative random variable T and the ordered pair $(\Pr\{T \geq c\}, E[T])$. By the definition of Ψ_c^* , we have $(\Pr\{T \geq c\}, E[T]) \in \Psi_c^* \subset R_f$, so that $(\Pr\{T \geq c\}, E[T]) \in R_f$. It then follows from the definition of R_f that $f(\Pr\{T \geq c\}, E[T]) \geq 0$.

Next, we show that if (12) holds for all nonnegative random variable T , then $\Psi_c^* \subset R_f$. Consider any $(p, m) \in \Psi_c^*$. Then $(p, m) = (\Pr\{T \geq c\}, E[T])$ for some nonnegative random variable T . Since by our assumption (12) holds for all nonnegative random variable T , we see that $(p, m) \in R_f$, and hence $\Psi_c^* \subset R_f$.

Combining the above, we conclude that $\Psi_c^* \subset R_f$ is a necessary and sufficient condition for the inequality set (12) to be valid. The theorem is proved. \square

We now consider a finite set of inequalities on $\Pr\{T \geq c\}$ and $E[T]$,

$$F = \{f_i(\Pr\{T \geq c\}, E[T]) \geq 0, 1 \leq i \leq k\}, \quad (13)$$

where $f_i : \mathbb{R}^2 \rightarrow \mathbb{R}$. We say that F is valid if $f_i(\Pr\{T \geq c\}, E[T]) \geq 0$ is valid for all i . Let

$$R_F = \{(p, m) \in \mathbb{R}^2 : f_i(p, m) \geq 0, 1 \leq i \leq m\}$$

be the region in \mathbb{R}^2 induced by F . The following is a corollary of Theorem 4.

Corollary 1. *The set of inequalities (13) is valid if and only if $\Psi_c^* \subset R_F$.*

Proof. Let

$$\tilde{f}(p, m) = \begin{cases} 1 & \text{if } f_i(p, m) \geq 0 \text{ for all } i \\ -1 & \text{otherwise.} \end{cases}$$

Then $\tilde{f}(p, m) \geq 0$ if and only if $f_i(p, m) \geq 0$ for all i . In other words, the set of inequalities (13) is valid if and only if $\tilde{f}(p, m) \geq 0$ is valid, and hence $R_F = R_{\tilde{f}}$. Then the corollary is proved by applying Theorem 4. \square

We end this section with an example for Corollary 1.

Example 1. *Let*

$$f_1 = p, \quad f_2 = 1 - p, \quad f_3 = m - cp,$$

and $F = \{f_i(\Pr\{T \geq c\}, E[T]) \geq 0, i = 1, 2, 3\}$. Then R_F becomes Ψ_c . Since $f_i(\Pr\{T \geq c\}, E[T]) \geq 0$ is valid for $i = 1, 2, 3$, so is the inequality set F . Then by Corollary 1, $\Psi_c^* \subset \Psi_F$, which is indeed the case.

4. Cauchy-Schwarz Inequality

The Cauchy-Schwarz inequality, which applies to a general inner product space, is among the most important inequalities in mathematics. For the purpose of this work, it suffices to confine our discussion to *real* inner product spaces.

Definition 1. *A real inner product space is a vector space V over \mathbb{R} together with an inner product $\langle \cdot, \cdot \rangle : V \times V \rightarrow \mathbb{R}$ that satisfies the following for any vectors $\mathbf{u}, \mathbf{v}, \mathbf{w} \in V$ and any scalars $a, b \in \mathbb{R}$:*

1. (Symmetry) $\langle \mathbf{u}, \mathbf{v} \rangle = \langle \mathbf{v}, \mathbf{u} \rangle$
2. (Linearity) $\langle a\mathbf{u} + b\mathbf{v}, \mathbf{w} \rangle = a\langle \mathbf{u}, \mathbf{w} \rangle + b\langle \mathbf{v}, \mathbf{w} \rangle$

3. (Positive-definiteness) $\langle \mathbf{u}, \mathbf{u} \rangle > 0$ if $\mathbf{u} \neq \mathbf{0}$.

The Cauchy-Schwarz inequality asserts that for any $\mathbf{u}, \mathbf{v} \in V$,

$$\langle \mathbf{u}, \mathbf{v} \rangle^2 \leq \langle \mathbf{u}, \mathbf{u} \rangle \langle \mathbf{v}, \mathbf{v} \rangle, \quad (14)$$

with equality if and only if \mathbf{u} and \mathbf{v} are linearly dependent. Here, for any pair of vectors \mathbf{u} and \mathbf{v} , the three quantities of interest are $\langle \mathbf{u}, \mathbf{u} \rangle$, $\langle \mathbf{v}, \mathbf{v} \rangle$, and $\langle \mathbf{u}, \mathbf{v} \rangle$, and we are interested in the relation among them. The Cauchy-Schwarz inequality gives a characterization of this relation. In this section, we discuss this inequality in the spirit of the discussion in the previous sections.

In view of the three quantities involved in the Cauchy-Schwarz inequality, namely $\langle \mathbf{u}, \mathbf{u} \rangle$, $\langle \mathbf{v}, \mathbf{v} \rangle$, and $\langle \mathbf{u}, \mathbf{v} \rangle$, we are motivated to consider the ordered triple $(\langle \mathbf{u}, \mathbf{u} \rangle, \langle \mathbf{v}, \mathbf{v} \rangle, \langle \mathbf{u}, \mathbf{v} \rangle)$ for any $\mathbf{u}, \mathbf{v} \in V$. For $(x, y, z) \in \mathbb{R}^3$, if $(x, y, z) = (\langle \mathbf{u}, \mathbf{u} \rangle, \langle \mathbf{v}, \mathbf{v} \rangle, \langle \mathbf{u}, \mathbf{v} \rangle)$ for some $\mathbf{u}, \mathbf{v} \in V$, then we say that (x, y, z) is *achievable*. We then define the *achievable region*

$$\Phi^* = \{(x, y, z) \in \mathbb{R}^3 : (x, y, z) \text{ is achievable}\}.$$

Also note that an ordered triple in \mathbb{R}^3 may be achievable by more than one pair of vectors, i.e.,

$$(x, y, z) = (\langle \mathbf{u}, \mathbf{u} \rangle, \langle \mathbf{v}, \mathbf{v} \rangle, \langle \mathbf{u}, \mathbf{v} \rangle) = (\langle \mathbf{u}', \mathbf{u}' \rangle, \langle \mathbf{v}', \mathbf{v}' \rangle, \langle \mathbf{u}', \mathbf{v}' \rangle)$$

where $(\mathbf{u}, \mathbf{v}) \neq (\mathbf{u}', \mathbf{v}')$.

Let V be a fixed inner product space. Consider a finite set of inequalities on the quantities $\langle \mathbf{u}, \mathbf{u} \rangle$, $\langle \mathbf{v}, \mathbf{v} \rangle$, and $\langle \mathbf{u}, \mathbf{v} \rangle$:

$$F = \{f_i(\langle \mathbf{u}, \mathbf{u} \rangle, \langle \mathbf{v}, \mathbf{v} \rangle, \langle \mathbf{u}, \mathbf{v} \rangle) \geq 0 : 1 \leq i \leq m\} \quad (15)$$

where $f_i : \mathbb{R}^3 \rightarrow \mathbb{R}$. For example, if $f_i(x, y, z) = xy - z^2$, then the i th inequality in (15) becomes (14), the Cauchy-Schwarz inequality. Let

$$R_F = \{(x, y, z) \in \mathbb{R}^3 : f_i(x, y, z) \geq 0, 1 \leq i \leq m\}$$

be the region in \mathbb{R}^3 induced by F .

If an inequality in (15) holds for all $\mathbf{u}, \mathbf{v} \in V$, we say that the inequality is *valid*. The Cauchy-Schwarz inequality is such an example. If the inequality in (15) is valid for all i , we say that the inequality set F is *valid*. The next theorem on the fundamental importance of the achievable region Φ^* follows directly from the discussion in Section 3, so the proof is omitted here.

Theorem 5. For any inner product space V , the inequality set (15) is valid if and only if $\Phi^* \subset R_F$.

Consider

$$f_1(x, y, z) = x \tag{16}$$

$$f_2(x, y, z) = y \tag{17}$$

$$f_3(x, y, z) = xy - z^2. \tag{18}$$

From (16) to (18), we obtain

$$f_1(\langle \mathbf{u}, \mathbf{u} \rangle, \langle \mathbf{v}, \mathbf{v} \rangle, \langle \mathbf{u}, \mathbf{v} \rangle) = \langle \mathbf{u}, \mathbf{u} \rangle \geq 0 \tag{19}$$

$$f_2(\langle \mathbf{u}, \mathbf{u} \rangle, \langle \mathbf{v}, \mathbf{v} \rangle, \langle \mathbf{u}, \mathbf{v} \rangle) = \langle \mathbf{v}, \mathbf{v} \rangle \geq 0 \tag{20}$$

$$f_3(\langle \mathbf{u}, \mathbf{u} \rangle, \langle \mathbf{v}, \mathbf{v} \rangle, \langle \mathbf{u}, \mathbf{v} \rangle) = \langle \mathbf{u}, \mathbf{u} \rangle \langle \mathbf{v}, \mathbf{v} \rangle - \langle \mathbf{u}, \mathbf{v} \rangle^2 \geq 0 \tag{21}$$

respectively, which hold for all $\mathbf{u}, \mathbf{v} \in V$. Note that (19) and (20) are implied by the positive-definiteness of an inner product space. From (19) to (21), we see that

$$f_i(\langle \mathbf{u}, \mathbf{u} \rangle, \langle \mathbf{v}, \mathbf{v} \rangle, \langle \mathbf{u}, \mathbf{v} \rangle) \geq 0$$

is valid for $i = 1, 2, 3$. Now define region

$$\Phi = \{(x, y, z) \in \mathbb{R}^3 : x, y \geq 0 \text{ and } z^2 \leq xy\}$$

which in fact is the region R_F with $F = \{f_i \geq 0, i = 1, 2, 3\}$. Thus by Theorem 5, we have $\Phi^* \subset \Phi$.

Ultimately, we are interested in obtaining a complete characterization of the achievable region Φ^* instead of just an outer bound on it. The question is whether Φ is indeed equal to Φ^* . If so, we say that the Cauchy-Schwarz inequality is *tight*. Otherwise, additional inequalities on the quantities $\langle \mathbf{u}, \mathbf{u} \rangle$, $\langle \mathbf{v}, \mathbf{v} \rangle$, and $\langle \mathbf{u}, \mathbf{v} \rangle$ are needed to completely characterize Φ^* . Such an inequality, if exists, would play the same fundamental role as the Cauchy-Schwarz inequality. The next theorem asserts that there exists no constraint on $\langle \mathbf{u}, \mathbf{u} \rangle$, $\langle \mathbf{v}, \mathbf{v} \rangle$, and $\langle \mathbf{u}, \mathbf{v} \rangle$ other than the Cauchy-Schwarz inequality and positive-definiteness if $\dim(V) \geq 2$, but is not so if $\dim(V) = 0, 1$. We also say that the Cauchy-Schwarz inequality is *sharp* if and only if $\dim(V) \geq 2$ (by regarding positive-definiteness as given).³

Theorem 6. For an inner product space V , $\Phi^* = \Phi$ if and only if $\dim(V) \geq 2$. For $\dim(V) = 0$, $\Phi^* = \{0\}$, and for $\dim(V) = 1$,

$$\Phi^* = \{(x, y, z) \in \mathbb{R}^3 : x, y \geq 0 \text{ and } z^2 = xy\}. \tag{22}$$

The following proposition is instrumental in proving Theorem 6.

Proposition 2. Let V be an inner product space. If $\dim(V) \geq 2$, then for any non-zero vector $\mathbf{u} \in V$, there exists a unit vector $\mathbf{w} \in V$ such that $\langle \mathbf{u}, \mathbf{w} \rangle = 0$.

Proof. Consider an inner product space V with $\dim(V) \geq 2$ and let \mathbf{u} be a non-zero vector in V . Since $\dim(V) \geq 2$, there exists another vector $\mathbf{t} \in V$ such that \mathbf{u} and \mathbf{t} are linearly independent. Let

$$\mathbf{v} = \mathbf{t} - \frac{\langle \mathbf{u}, \mathbf{t} \rangle}{\langle \mathbf{u}, \mathbf{u} \rangle} \mathbf{u}.$$

Note that $\langle \mathbf{u}, \mathbf{u} \rangle > 0$ since $\mathbf{u} \neq 0$. Then

$$\begin{aligned} \langle \mathbf{u}, \mathbf{v} \rangle &= \left\langle \mathbf{u}, \mathbf{t} - \frac{\langle \mathbf{u}, \mathbf{t} \rangle}{\langle \mathbf{u}, \mathbf{u} \rangle} \mathbf{u} \right\rangle \\ &= \langle \mathbf{u}, \mathbf{t} \rangle - \frac{\langle \mathbf{u}, \mathbf{t} \rangle \langle \mathbf{u}, \mathbf{u} \rangle}{\langle \mathbf{u}, \mathbf{u} \rangle} \\ &= 0, \end{aligned}$$

where the last step is justified because $\langle \mathbf{u}, \mathbf{u} \rangle \neq 0$. Also, we have

$$\begin{aligned} \langle \mathbf{v}, \mathbf{v} \rangle &= \left\langle \mathbf{t} - \frac{\langle \mathbf{u}, \mathbf{t} \rangle}{\langle \mathbf{u}, \mathbf{u} \rangle} \mathbf{u}, \mathbf{t} - \frac{\langle \mathbf{u}, \mathbf{t} \rangle}{\langle \mathbf{u}, \mathbf{u} \rangle} \mathbf{u} \right\rangle \\ &= \langle \mathbf{t}, \mathbf{t} \rangle - 2 \frac{\langle \mathbf{u}, \mathbf{t} \rangle^2}{\langle \mathbf{u}, \mathbf{u} \rangle} + \frac{\langle \mathbf{u}, \mathbf{t} \rangle^2 \langle \mathbf{u}, \mathbf{u} \rangle}{\langle \mathbf{u}, \mathbf{u} \rangle^2} \\ &= \langle \mathbf{t}, \mathbf{t} \rangle - \frac{\langle \mathbf{u}, \mathbf{t} \rangle^2}{\langle \mathbf{u}, \mathbf{u} \rangle} \\ &\geq 0, \end{aligned}$$

where the inequality above follows from the Cauchy-Schwarz inequality (14) which is tight if and only if \mathbf{u} and \mathbf{t} are linearly dependent. Since the latter does not hold by our assumption, we conclude that $\langle \mathbf{v}, \mathbf{v} \rangle > 0$, i.e., \mathbf{v} is not the zero vector. Finally, the proposition is proved by letting

$$\mathbf{w} = \frac{\mathbf{v}}{\sqrt{\langle \mathbf{v}, \mathbf{v} \rangle}},$$

so that

$$\langle \mathbf{u}, \mathbf{w} \rangle = \frac{\langle \mathbf{u}, \mathbf{v} \rangle}{\sqrt{\langle \mathbf{v}, \mathbf{v} \rangle}} = 0$$

and

$$\sqrt{\langle \mathbf{w}, \mathbf{w} \rangle} = \sqrt{\frac{\langle \mathbf{v}, \mathbf{v} \rangle}{\langle \mathbf{v}, \mathbf{v} \rangle}} = 1.$$

□

Proof of Theorem 6. For the case $\dim(V) = 0$, since $V = \{0\}$, we have $\Phi^* = \{0\} \subsetneq \Phi$.

For the case $\dim(V) = 1$, since for any two vectors $\mathbf{u}, \mathbf{v} \in V$, one of them is always a scalar multiple of the other, the inequality (14) is always satisfied with equality. This implies that

$$\Phi^* \subset \{(x, y, z) \in \mathbb{R}^3 : x, y \geq 0 \text{ and } z^2 = xy\}, \quad (23)$$

which is a proper subset of Φ . To prove that

$$\{(x, y, z) \in \mathbb{R}^3 : x, y \geq 0 \text{ and } z^2 = xy\} \subset \Phi^*,$$

it suffices to show that every (x, y, z) satisfying $x, y \geq 0$ and $z^2 = xy$, there exist $\mathbf{u}, \mathbf{v} \in V$ such that

$$(x, y, z) = (\langle \mathbf{u}, \mathbf{u} \rangle, \langle \mathbf{v}, \mathbf{v} \rangle, \langle \mathbf{u}, \mathbf{v} \rangle). \quad (24)$$

We first consider the case that either $x = 0$ or $y = 0$. If $x = 0$, then $z^2 \leq xy$ becomes $z^2 \leq 0$, which implies $z = 0$. Then (24) is satisfied by letting $\mathbf{u} = 0$ and $\mathbf{v} \in V$ such that $\langle \mathbf{v}, \mathbf{v} \rangle = y$. Likewise if $y = 0$.

Now consider the case that $x, y > 0$. Then $z^2 = xy > 0$, which implies that $z \neq 0$. Let $\mathbf{u} \in V$ such that

$$\langle \mathbf{u}, \mathbf{u} \rangle = x. \quad (25)$$

We need to consider two cases for z , namely $z > 0$ and $z < 0$. First consider the case $z > 0$. Together with $z^2 = xy$, we have $z = \sqrt{xy}$. Let $b = \sqrt{y/x}$, so that $b^2x = y$. Let $\mathbf{v} = b\mathbf{u}$. Then

$$\langle \mathbf{v}, \mathbf{v} \rangle = \langle b\mathbf{u}, b\mathbf{u} \rangle = b^2 \langle \mathbf{u}, \mathbf{u} \rangle = b^2x = y, \quad (26)$$

and

$$\langle \mathbf{u}, \mathbf{v} \rangle = \langle \mathbf{u}, b\mathbf{u} \rangle = b \langle \mathbf{u}, \mathbf{u} \rangle = bx = \sqrt{xy} = z. \quad (27)$$

Then we see from (25) to (27) that (24) is satisfied. For the case $z < 0$, we have $z = -\sqrt{xy}$. Then we let $b = -\sqrt{y/x}$ and repeat the above steps to show that (24) is again satisfied. Therefore, we have proved (23) and hence (22).

Now consider the case $\dim(V) \geq 2$. It suffices to show that for any $(x, y, z) \in \Phi$, there exist $\mathbf{u}, \mathbf{v} \in V$ such that (24) is satisfied. Then $(x, y, z) \in \Phi^*$, showing that $\Phi \subset \Phi^*$.

Consider any $(x, y, z) \in \Phi$. Let $\mathbf{u} \in V$ such that $\langle \mathbf{u}, \mathbf{u} \rangle = x$. We seek $\mathbf{v} \in V$ such that

$$\langle \mathbf{v}, \mathbf{v} \rangle = y \quad (28)$$

and

$$\langle \mathbf{u}, \mathbf{v} \rangle = z. \quad (29)$$

We first consider the case that either $x = 0$ or $y = 0$, which can be proved in exactly the same way as we have proved the case for $\dim(V) = 1$. Now consider the case that $x, y > 0$, and choose any $\mathbf{u} \in V$ such

that $\langle \mathbf{u}, \mathbf{u} \rangle = x$. From Proposition 2, there exists a unit vector $\mathbf{w} \in V$ such that $\langle \mathbf{u}, \mathbf{w} \rangle = 0$. Let

$$\mathbf{v} = \frac{z}{x}\mathbf{u} + \sqrt{y - \frac{z^2}{x}}\mathbf{w}.$$

Note that the quantity inside the square root is nonnegative because $z^2 \leq xy$. Since $\langle \mathbf{u}, \mathbf{w} \rangle = 0$, we have

$$\langle \mathbf{v}, \mathbf{v} \rangle = \frac{z^2}{x^2}\langle \mathbf{u}, \mathbf{u} \rangle + \left(y - \frac{z^2}{x}\right) = \frac{z^2}{x^2}x + \left(y - \frac{z^2}{x}\right) = \frac{z^2}{x} + \left(y - \frac{z^2}{x}\right) = y,$$

and

$$\langle \mathbf{u}, \mathbf{v} \rangle = \left\langle \mathbf{u}, \frac{z}{x}\mathbf{u} \right\rangle = \frac{z}{x}\langle \mathbf{u}, \mathbf{u} \rangle = z.$$

Thus (28) and (29) are satisfied. The theorem is proved. \square

Remarks

1. From the proof of Theorem 6, we see that when $\dim(V) = 1$, Φ^* is exactly equal to the boundary of Φ .
2. From Theorem 6, we see that Φ^* for $\dim(V) = 0$ or $\dim(V) = 1$ is a subset of Φ^* for $\dim(V) \geq 2$. Therefore, there exists no inequality on the quantities $\langle \mathbf{u}, \mathbf{u} \rangle$, $\langle \mathbf{v}, \mathbf{v} \rangle$, and $\langle \mathbf{u}, \mathbf{v} \rangle$ that holds for all inner product space V (regardless of the value of $\dim(V)$) which is not implied by the Cauchy-Schwarz inequality.

To end this section, we argue that the Cauchy-Schwarz inequality can be regarded as sharp even when $\dim(V) = 0$ or 1 if we take explicit consideration of the dimension of the vector space. Specifically,

- If $\dim(V) = 0$, then for any $\mathbf{u}, \mathbf{v} \in V$, we have $\mathbf{u} = \mathbf{v} = \mathbf{0}$, so that $(\langle \mathbf{u}, \mathbf{u} \rangle, \langle \mathbf{v}, \mathbf{v} \rangle, \langle \mathbf{u}, \mathbf{v} \rangle) = (0, 0, 0)$.

With this additional constraint, we can refined Φ to

$$\Phi_0 = \Phi \cap \{0\} = \{0\},$$

which is equal to Φ^* .

- If $\dim(V) = 1$, for any $\mathbf{u}, \mathbf{v} \in V$, one of them is always a scalar multiple of the other, and the inequality (14) is always satisfied with equality. Then by imposing this additional constraint, we can refine Φ to

$$\Phi_1 = \Phi \cap \{(x, y, z) \in \mathbb{R}^3 : z^2 = xy\},$$

which again is equal to Φ^* .

5. Entropy Inequalities

In the last section, we use the Cauchy-Schwarz inequality to illustrate how a geometrical formulation can potentially lead to interesting results. In this section, we apply the same formality to inequalities on the Shannon entropy, which has led to very fruitful and unexpected results in the past over two decades. This section is a brief exposition of this subject. The reader is referred to [4], Chs. 13-15] and [5][6] for more in-depth discussions.⁴

In this section, all random variables are discrete. The Shannon entropy (or simply entropy when there is no ambiguity) for a random variable X with probability mass function $p(x)$ is defined as

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

where \mathcal{S}_X denotes the support of X . For a pair of jointly distributed random variables X and Y with probability mass function $p(x, y)$, the entropy is defined as

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

where \mathcal{S}_{XY} denotes the support of $p(x, y)$. The entropy for a finite number of random variables is defined likewise. The entropy for two or more random variables is often called a *joint entropy*, although the distinction between entropy and joint entropy is unnecessary.

In information theory (see [7][8][4]), entropy is the fundamental measure of information. In addition to entropy, the following quantities are defined:

<i>Mutual Information</i>	$I(X; Y)$ $= H(X) + H(Y) - H(X, Y)$
<i>Conditional Entropy</i>	$H(X Y)$ $= H(X, Y) - H(Y)$
<i>Conditional Mutual Information</i>	$I(X; Y Z)$ $= H(X, Z) + H(Y, Z) - H(X, Y, Z) - H(Z)$

These quantities, collectively called *Shannon's information measures*, are used extensively in coding theorems in information theory problems.

In this section, inequalities on Shannon's information measures are discussed. These inequalities are the main tool for proving converse coding theorems, which establish that for a particular communication problem, no coding scheme exists if certain conditions are not satisfied. In other words, these inequalities establishes the "impossibilities" in information theory, and they are sometimes referred to as the "laws of information theory" [9].

As we see from the above, all Shannon's information measures can be expressed as a linear combinations of entropies. Therefore, inequalities on Shannon's information measures can be written as inequalities on entropies. For this reason, they are referred to as *entropy inequalities*.

In this section, we will not focus on the application of entropy inequalities in proving converse coding theorems. Rather, we will focus on these inequalities themselves. Like what we have done in the previous sections, we first introduce a geometrical framework for entropy inequalities.

Let $[n] = \{1, \dots, n\}$, $\mathbf{N} = 2^{[n]}$, and $\bar{\mathbf{N}} = \mathbf{N} \setminus \{\emptyset\}$. Let $\Theta = \{X_i, i \in [n]\}$ be a collection of n discrete random variables. Associated with any collection of n random variables are $k := 2^n - 1$ joint entropies. For $\alpha \in \mathbf{N}$, write $X_\alpha = (X_i, i \in \alpha)$, with the convention that X_\emptyset is a constant. For example, $X_{\{1,2,3\}}$, or simply X_{123} , denotes (X_1, X_2, X_3) . For a collection Θ of n random variables, define the set function $H_\Theta : \mathbf{N} \rightarrow \mathbb{R}$ by

$$H_\Theta(\alpha) = H(X_\alpha), \quad \alpha \in \mathbf{N},$$

with $H_\Theta(\emptyset) = 0$ because X_\emptyset is a constant. H_Θ is called the *entropy function* of Θ .

Let \mathcal{H}_n denote \mathbb{R}^k , the k -dimensional Euclidean space, with the coordinates labeled by $h_\alpha, \alpha \in \bar{\mathbf{N}}$. We call \mathcal{H}_n the *entropy space* for n random variables. As an example, for $n = 3$, the coordinates of \mathcal{H}_3 are labelled by

$$h_1, h_2, h_3, h_{12}, h_{13}, h_{23}, h_{123},$$

where h_{123} denotes $h_{\{1,2,3\}}$, etc. Then for each collection Θ of n random variables, H_Θ can be represented by a column vector $\mathbf{h}^\Theta \in \mathcal{H}_n$, called the *entropy vector* of Θ , whose component corresponding to α is equal to $H_\Theta(\alpha)$ for all $\alpha \in \bar{\mathbf{N}}$. On the other hand, a column vector $\mathbf{h} \in \mathcal{H}_n$ is called *entropic*⁵ if it is equal to the entropy vector \mathbf{h}^Θ of some collection Θ of n random variables.

5.1. Unconstrained and Constrained Entropy Inequalities

Like what we have done for the Markov inequality and the Cauchy-Schwarz inequality, we are motivated to define the region

$$\Gamma_n^* = \{\mathbf{h} \in \mathbb{R}^k : \mathbf{h} \text{ is entropic}\}.$$

The region Γ_n^* is referred to as the region of entropy vectors.

An entropy inequality $f(\mathbf{h}^\Theta) \geq 0$, where $f : \mathbb{R}^k \rightarrow \mathbb{R}$, is *valid* if it holds for all collection Θ of n random variables. For example, the inequality

$$H(X_1) + H(X_2) \geq H(X_1, X_2),$$

or

$$I(X_1; X_2) \geq 0,$$

is valid because it holds for any random variables X_1 and X_2 ; this will be further discussed in Section 5.2.

In the sequel, let n be fixed. The following proposition is analogous to Proposition 4. Its proof is omitted.

Proposition 3. *A set of entropy inequalities $\{f_i(\mathbf{h}^\ominus) \geq 0, 1 \leq i \leq m\}$ is valid if and only if*

$$\Gamma_n^* \subseteq \{\mathbf{h} \in \mathcal{H}_n : f_i(\mathbf{h}) \geq 0, 1 \leq i \leq m\}.$$

In information theory, we very often deal with entropy inequalities with certain constraints on the joint distribution for the random variables involved. These are called constrained entropy inequalities, and the constraints on the joint distribution can usually be expressed as linear constraints on the entropies. In the sequel, we always assume that the constraints on the entropies are of this form. The following are some examples:

1. X_1 is a function of X_2 if and only if $H(X_1|X_2) = 0$.
2. X_1 and X_2 are independent conditioning on X_3 if and only if $I(X_1; X_2|X_3) = 0$.
3. The Markov chain $X_1 \leftrightarrow X_2 \leftrightarrow X_3 \leftrightarrow X_4$ holds if and only if
$$\begin{cases} I(X_1; X_3 | X_2) = 0 \\ I(X_1, X_2; X_4 | X_3) = 0 \end{cases}$$
4. Three random variables $X_1, X_2,$ and X_3 are mutually independent if and only if

$$H(X_1, X_2, X_3) = H(X_1) + H(X_2) + H(X_3). \quad (30)$$

It is not difficult to show that (30) is equivalent to

$$\begin{cases} I(X_1; X_2) = 0 \\ I(X_2; X_3|X_1) = 0 \\ I(X_1; X_3|X_2) = 0. \end{cases}$$

Suppose there are q constraints on the entropies given by

$$Q\mathbf{h} = 0,$$

where Q is a $q \times k$ matrix. Without loss of generality, we can assume that these q constraints are linearly independent, so that Q is full row rank. Let

$$\Phi = \{\mathbf{h} \in \mathcal{H}_n : Q\mathbf{h} = 0\}. \quad (31)$$

In other words, the q constraints confine \mathbf{h} to a linear subspace Φ in the entropy space. The following is the constrained version of Proposition 3.

Proposition 4. *A set of entropy inequalities $\{f_i(\mathbf{h}^\ominus) \geq 0, 1 \leq i \leq m\}$ is valid under the constraint Φ if and only if*

$$(\Gamma_n^* \cap \Phi) \subset \{\mathbf{h} \in \mathcal{H}_n : f_i(\mathbf{h}) \geq 0, 1 \leq i \leq m\}.$$

We will refer to the inequalities in Proposition 3 as (unconstrained) entropy inequalities, and the inequalities in Proposition 4 as constrained entropy inequalities. Note that we can let $\Phi = \mathcal{H}_n$ when there is no constraint on the entropies. In this sense, an unconstrained entropy inequality is a special case of a constrained entropy inequality.

5.2. Shannon-Type Inequalities

It is well known in information theory that all Shannon's information measures are nonnegative. This set of inequalities is collectively called the *basic inequalities* in information theory. Specifically, these are inequalities of the form

1. $H(X_\alpha) \geq 0$,
2. $I(X_\alpha; X_\beta) \geq 0$,
3. $H(X_\alpha | X_\gamma) \geq 0$,
4. $I(X_\alpha; X_\beta | X_\gamma) \geq 0$,

where α , β , and γ are disjoint subsets of \mathbf{N} . On the other hand, the entropy function satisfies the polymatroidal axioms ^[10]: For any $\delta, \sigma \subset \mathbf{N}$,

1. $H_\Theta(\emptyset) = 0$;
2. $H_\Theta(\delta) \leq H_\Theta(\sigma)$ if $\delta \subset \sigma$;
3. $H_\Theta(\delta) + H_\Theta(\sigma) \geq H_\Theta(\delta \cup \sigma) + H_\Theta(\delta \cap \sigma)$.

It can be shown that the basic inequalities and polymatroid axioms are equivalent (see ^[4][Appendix 14.A]).

The basic inequalities, expressed in terms of the entropies, are linear inequalities in \mathcal{H}_n . Denote this set of inequalities by $G\mathbf{h} \geq 0$, where G is an $m \times k$ matrix, and define

$$\Gamma_n = \{\mathbf{h} \in \mathcal{H}_n : G\mathbf{h} \geq 0\}.$$

Since the basic inequalities always hold, we see from Proposition 3 that $\Gamma_n^* \subset \Gamma_n$.

Shannon-type inequalities are entropy inequalities that are implied by the basic inequalities. Specifically, an entropy inequality $f(\mathbf{h}) \geq 0$ is a Shannon-type inequality if and only if

$$\Gamma_n \subset \{\mathbf{h} \in \mathcal{H}_n : f(\mathbf{h}) \geq 0\}.$$

More generally, under the linear constraint Φ (cf. (31)), $f(\mathbf{h}) \geq 0$ is a Shannon-type inequality if and only if

$$(\Gamma_n \cap \Phi) \subset \{\mathbf{h} \in \mathcal{H}_n : f(\mathbf{h}) \geq 0\}.$$

Since $\Gamma_n^* \subset \Gamma_n$, it follows that

$$(\Gamma_n^* \cap \Phi) \subset \{\mathbf{h} \in \mathcal{H}_n : f(\mathbf{h}) \geq 0\},$$

which implies that a Shannon-type inequality is valid.

As mentioned earlier in this section, entropy inequalities are the main tool for proving converse coding theorems in information theory. In fact, Shannon-type inequalities had been all the entropy inequalities that were known until the discovery of *non-Shannon-type* inequalities in the late 1990s.

Since Γ_n is a polyhedral cone, verification of a *linear* Shannon-type inequality can be formulated as a linear programming problem [2]. ITIP [11] was the first software developed for this purpose, which runs on MATLAB. Subsequently, variants of ITIP with different additional features have been developed. AITIP [12] can produce a human-readable proof and suggest counterexamples when the inequality to be verified is not Shannon-type. PSITIP [13] can render proofs for converse coding theorems in network information theory [14]. See [15] for a list of related software. Recently, a symbolic approach to the problem that can drastically speed up the computation has been developed [16][17]. For a general discussion on machine-proving of entropy inequalities, we refer the reader to the tutorial paper [18].

5.3. Beyond Shannon-Type Inequalities

To our knowledge, [9] was the first work in the literature that explicitly asked whether there exists any constraint on the entropy function other than the polymatroidal axioms. The same question was raised in [19] in a somewhat different form. With the geometrical formulation for entropy inequalities described at the beginning of this section, it became reasonable to conjecture the existence of constraints on the entropy function beyond Shannon-type inequalities, because it is not readily provable that $\Gamma_n^* = \Gamma_n$.

Before diving deeper into this subject, we first note that

1. $\Gamma_2^* = \Gamma_2$;
2. $\Gamma_3^* \neq \Gamma_3$ but $\overline{\Gamma_3^*} = \Gamma_3$, where $\overline{\Gamma_3^*}$ denotes the closure of Γ_3^* ;

While it is straightforward to show that $\Gamma_2^* = \Gamma_2$, the problem is already nontrivial for $n = 3$. Specifically, along an extreme direction of Γ_3 , only certain discrete points are entropic, making Γ_3^* not closed. However, upon taking the closure of Γ_3^* , we obtain Γ_3 .

For $n \geq 4$, the set Γ_n^* is very complex and characterization of Γ_n^* remains an open problem. Nevertheless, the following general properties of Γ_n^* are known:

1. $\overline{\Gamma_n^*}$ is a convex cone;
2. $\text{int}(\overline{\Gamma_n^*}) \subset \Gamma_n^*$, where $\text{int}(\cdot)$ denotes the interior of a set ^[20].

Here, $\text{int}(\overline{\Gamma_n^*}) \subset \Gamma_n^*$ means that the difference between Γ_n^* and $\overline{\Gamma_n^*}$ can only be on the boundary. As discussed, Γ_3^* and $\overline{\Gamma_3^*}$ differ on an extreme direction of $\overline{\Gamma_3^*}$ ($= \Gamma_3$), which is on the boundary of $\overline{\Gamma_3^*}$.

For 4 random variables, the following constrained entropy inequality was proved ^[21]: If

$$I(X_1; X_2) = I(X_1; X_2|X_3) = 0, \tag{32}$$

then

$$I(X_3; X_4) \leq I(X_3; X_4|X_1) + I(X_3; X_4|X_2). \tag{33}$$

This inequality, referred to in the literature as ZY97, cannot be proved by ITIP and hence is a non-Shannon-type inequality.

It was discussed earlier that along an extreme direction of Γ_3 , only certain discrete points are entropic, while the rest are non-entropic. In the above, the constraints in (32) together with Γ_4 define a 13-dimensional face⁶ of Γ_4 , and ZY97 asserts that a region on this face is not entropic. However, it is still unclear whether $\overline{\Gamma_4^*}$ is equal to Γ_4 . Shortly after, the following unconstrained non-Shannon-type entropy inequality was discovered ^[22]: For any random variables X_1, X_2, X_3 , and X_4 ,

$$2I(X_3; X_4) \leq I(X_1; X_2) + I(X_1; X_3, X_4) + 3I(X_3; X_4|X_1) + I(X_3; X_4|X_2). \tag{34}$$

This inequality, referred to in the literature as ZY98 or the *Zhang-Yeung inequality*, shows that $\overline{\Gamma_4^*}$ is a proper subset of Γ_4 . See Fig. 4 for an illustration.

Subsequently, many non-Shannon-type inequalities for four or more random variables were discovered ^{[23][24][20][25][26]}. In particular, the existence of an infinite class of unconstrained non-Shannon-type inequalities for four random variables was proved, implying that Γ_4^* (and more generally Γ_n^*) is not a pyramid ^[20]. See ^[27] for a unifying discussion.

The above are the efforts on characterizing $\overline{\Gamma}_n^*$, in particular $\overline{\Gamma}_4^*$, which remains an open problem. Throughout the years, there have been efforts on characterizing Γ_4^* , which is even more difficult. Notable works along this line include [28][29][30][31][32]. The connection with Γ_n^* with conditional independence of random variables will be discussed in Section 5.4.2.

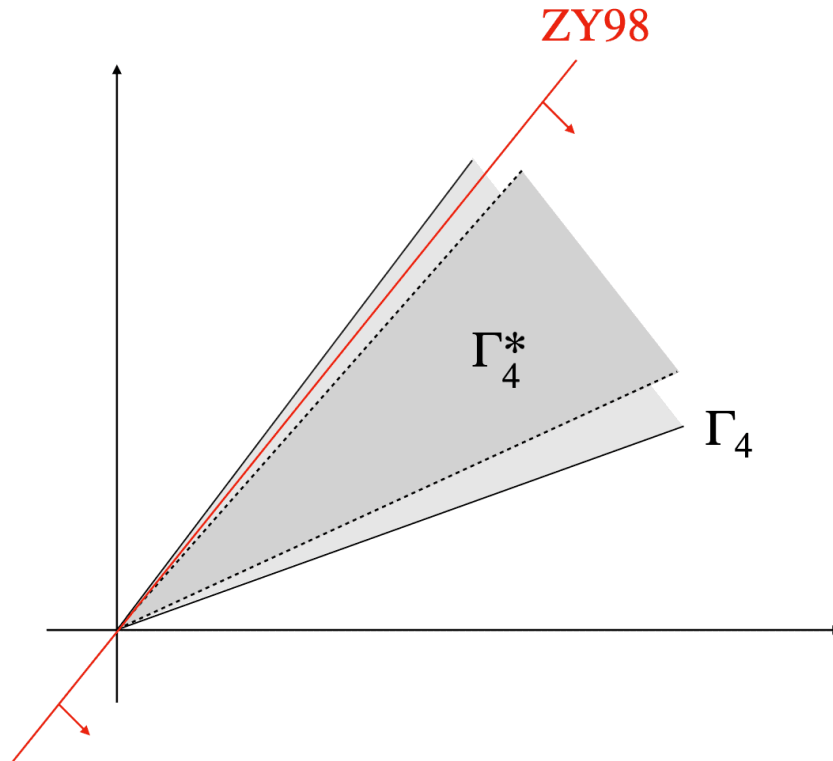


Figure 4. An illustration of the non-Shannon-type inequality ZY98.

5.4. Connections with Other Fields

The study of entropy inequalities, more specifically characterization of the region Γ_n^* , was shown to be intimately related to a distributed coding problem inspired by satellite communication [33]. This line of research was subsequently developed into the theory of network coding [34][35]. In the meantime, intimate relations between this subject and different branches of mathematics and physics were established. In particular, the non-Shannon-type inequalities for entropy induce corresponding inequalities for finite groups, Kolmogorov complexity, and positive semi-definite matrices. In this

section, we give a high-level introduction of these developments. For a comprehensive treatment of the topic, we refer the readers to [5][6].

5.4.1. Network Coding

In network communication, to send information from a source node s to a destination node t , the predominant existing method is *routing*, namely that data packets are routed from node s to node t through the intermediate nodes in its original form. Network coding theory [34] refutes the folklore that routing alone can achieve the network capacity. Rather, coding at the network nodes, referred to as “network coding”, is in general required. As routing a data packet from an input to an output of a network node can be regarded as applying the identity map to the data packet, routing is a special case of network coding.

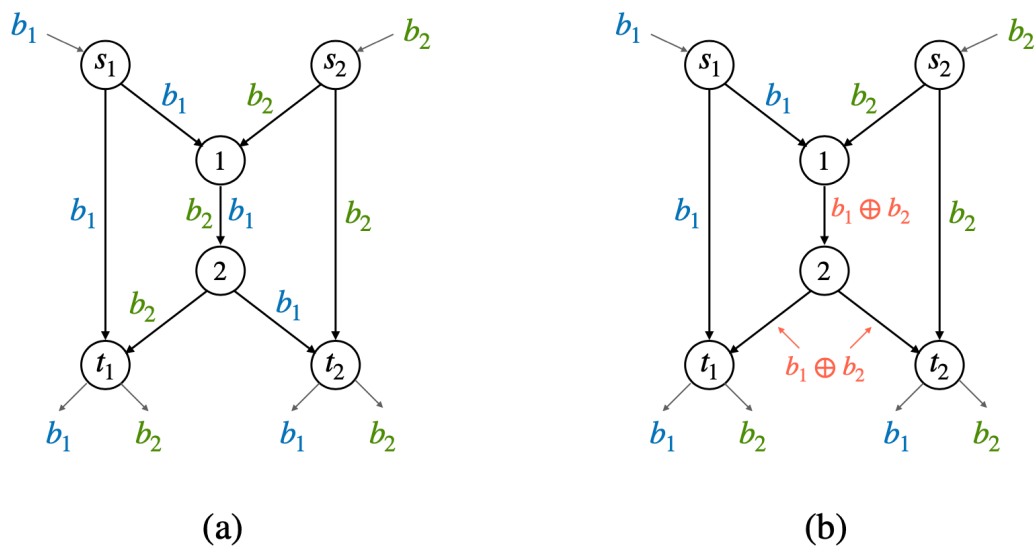


Figure 5. The butterfly network that illustrates the advantage of network coding.

The advantage of network coding can be illustrated by a simple example called the *butterfly network*, represented by the directed graph in Figure 5. Here, a directed edge (i, j) represents a communication channel from node i to node j . A bit b_i is generated at source node $s_i, i = 1, 2$, and the bits b_1 and b_2 are to be multicast⁷ to two destination nodes t_1 and t_2 . Figure 5(a) shows a routing solution, in which both b_1 and b_2 need to be transmitted on channel $(1, 2)$.

If only one bit can be transmitted on each channel, then there exists no routing solution for this multicast problem. However, if an intermediate node can apply computation to the incoming bits instead of just routing them through, then a solution can be obtained as in Figure 5(b). Here at node 1, the received bits b_1 and b_2 are combined into a new bit $b_1 \oplus b_2$, where ‘ \oplus ’ denotes binary addition or exclusive-or (XOR). This operation is referred to as *network coding*. The bit $b_1 \oplus b_2$ is then transmitted to node 2, where two copies of the bit are sent to the destination nodes t_1 and t_2 , respectively. At node t_1 , a copy of the bit b_1 is received directly from node s_1 , while the bit b_2 can be decoded by adding the two received bits:

$$b_1 \oplus (b_1 \oplus b_2) = (b_1 \oplus b_1) \oplus b_2 = 0 \oplus b_2 = b_2.$$

Similarly, the bits b_1 and b_2 can be decoded at node t_2 .

Among the vast literature of network coding, [33], [36], and [35] are directly related to the discussions in this section. In a nutshell, these works give a complete characterization of the *capacity region*⁸ of the general network coding problem on an acyclic network in terms of Γ^* , the region of entropy vectors. Here, we omit the subscript n in Γ^* because the exact number of random variables involved depends on the setup of the specific network coding problem. Evidently, this characterization is implicit because the complete characterization of Γ^* is still open. Nevertheless, an outer (inner) bound on Γ^* directly induces an outer (inner) bound on the capacity region of the network coding problem. For a comprehensive discussion on this topic, we refer the reader to [4], Ch. 21[35].

5.4.2. Probability Theory

We use $X_\alpha \perp X_\beta \mid X_\gamma$ to denote the conditional independency (CI)

$$X_\alpha \text{ and } X_\beta \text{ are conditionally independent given } X_\gamma,$$

where α , β , and γ are assumed to be disjoint subsets of $[n]$. When $\gamma = \emptyset$, $X_\alpha \perp X_\beta \mid X_\gamma$ becomes an unconditional independency which we regard as a special case of a conditional independency.

In probability theory, we are often given a set of CI’s and we need to determine whether another given CI is logically implied. We refer to this problem as the *implication problem*, which is one of the most basic problems in probability theory. For example, we want to know whether

$$\left. \begin{array}{l} X_1 \perp X_3 \mid X_2 \\ X_1 \perp X_2 \end{array} \right\} \Rightarrow X_1 \perp X_3.$$

This is not difficult to prove. However, the general implication problem is extremely difficult, and it has been solved only up to four random variables [37].

We now explain the relation between the implication problem and the region Γ_n^* . A CI involving random variables X_1, X_2, \dots, X_n has the form

$$X_\alpha \perp X_\beta \mid X_\gamma, \quad (35)$$

where α , β , and γ are disjoint subsets of $[n]$. Denote this generic CI by K . From the discussion in Section 5.1, K is equivalent to $I(X_\alpha; X_\beta \mid X_\gamma) = 0$, i.e., setting the basic inequality $I(X_\alpha; X_\beta \mid X_\gamma) \geq 0$ to equality. Furthermore, since $I(X_\alpha; X_\beta \mid X_\gamma) = 0$ is equivalent to

$$H(X_{\alpha \cup \gamma}) + H(X_{\beta \cup \gamma}) - H(X_{\alpha \cup \beta \cup \gamma}) - H(X_\gamma) = 0, \quad (36)$$

the CI K corresponds to the following hyperplane in \mathcal{H}_n :

$$\mathcal{E}(K) := \{\mathbf{h} \in \mathcal{H}_n : h_{\alpha \cup \gamma} + h_{\beta \cup \gamma} - h_{\alpha \cup \beta \cup \gamma} - h_\gamma = 0\}. \quad (37)$$

Since the region Γ_n is defined by the basic inequalities (with $I(X_\alpha; X_\beta \mid X_\gamma) \geq 0$ being one), $\mathcal{E}(K) \cap \Gamma_n$ is a face of Γ_n .

Let $\Pi = \{K_l\}$ be a collection of CIs, and we want to determine whether Π implies a given CI K . This would be the case if and only if the following is true:

$$\text{For all } \mathbf{h} \in \Gamma_n^*, \text{ if } \mathbf{h} \in \cap_l \mathcal{E}(K_l), \text{ then } \mathbf{h} \in \mathcal{E}(K).$$

Equivalently,

$$\Pi \text{ implies } K \quad \text{if and only if} \quad [\cap_l \mathcal{E}(K_l)] \cap \Gamma_n^* \subseteq \mathcal{E}(K).$$

Therefore, the implication problem can be solved if Γ_n^* can be characterized. Since $\Gamma_n^* \subseteq \Gamma_n$, in the above, $[\cap_l \mathcal{E}(K_l)] \cap \Gamma_n^*$ can be rewritten as

$$[\cap_l \mathcal{E}(K_l)] \cap \Gamma_n^* \cap \Gamma_n = \cap_l [(\mathcal{E}(K_l) \cap \Gamma_n) \cap \Gamma_n^*].$$

As discussed, $\mathcal{E}(K_l) \cap \Gamma_n$ is a face of Γ_n . In other words, the implication problem can be solved by characterizing Γ_n^* , in particular characterizing Γ_n^* on the faces of Γ_n . Therefore, to tackle the implication problem, it is not sufficient just to characterize $\overline{\Gamma_n^*}$.

Hence, the region Γ_n^* is not only of fundamental importance in information theory, but is also of fundamental importance in probability theory. For a more general discussion of this topic, we refer the reader to the series of papers [38][39][37].

5.4.3. Group Theory

Let X_1 and X_2 be any two random variables. Then

$$H(X_1) + H(X_2) \geq H(X_1, X_2), \quad (38)$$

which is equivalent to the basic inequality

$$I(X_1; X_2) \geq 0. \quad (39)$$

Let G be any finite group and G_1 and G_2 be subgroups of G . It is well known in group theory⁹ that

$$|G||G_1 \cap G_2| \geq |G_1||G_2|, \quad (40)$$

where $|G|$ denotes the *order* of G and $G_1 \cap G_2$ denotes the *intersection* of G_1 and G_2 ($G_1 \cap G_2$ is also a subgroup of G). By rearranging the terms, the above inequality can be written as

$$\log \frac{|G|}{|G_1|} + \log \frac{|G|}{|G_2|} \geq \log \frac{|G|}{|G_1 \cap G_2|}. \quad (41)$$

By comparing (38) and (41), one can easily identify the one-to-one correspondence between the forms of these two inequalities, namely that X_i corresponds to G_i , $i = 1, 2$, and (X_1, X_2) corresponds to $G_1 \cap G_2$. While (38) is true for any pair of random variables X_1 and X_2 , (41) is true for any finite group G and subgroups G_1 and G_2 . As a further example, from the entropy inequality

$$H(X_1, X_3) + H(X_2, X_3) \geq H(X_1, X_2, X_3) + H(X_3) \quad (42)$$

which is equivalent to $I(X_1; X_2|X_3) \geq 0$, we can obtain the group inequality

$$\log \frac{|G|}{|G_1 \cap G_3|} + \log \frac{|G|}{|G_2 \cap G_3|} \geq \log \frac{|G|}{|G_1 \cap G_2 \cap G_3|} + \log \frac{|G|}{|G_3|} \quad (43)$$

that holds for all finite group $|G|$ and subgroups G_1, G_2 , and G_3 .

This one-to-one correspondence can be extended to any random variables X_1, X_2, \dots, X_n and any finite group G and its subgroups G_1, G_2, \dots, G_n . For example, consider the non-Shannon-type inequality ZY98 which can be written in terms of joint entropies as follows:

$$\left. \begin{aligned} H(X_1) + H(X_1, X_2) + 2H(X_3) \\ + 2H(X_4) + 4H(X_1, X_3, X_4) \\ + H(X_2, X_3, X_4) \end{aligned} \right\} \leq \left\{ \begin{aligned} 3H(X_1, X_3) + 3H(X_1, X_4) \\ + 3H(X_3, X_4) + H(X_2, X_3) \cdot \\ + H(X_2, X_4) \end{aligned} \right.$$

This entropy inequality, which holds for all random variables X_1, X_2, X_3 and X_4 , corresponds to the group inequality

$$\left. \begin{aligned} &|G_1 \cap G_3|^3 |G_1 \cap G_4|^3 \\ &\cdot |G_3 \cap G_4|^3 |G_2 \cap G_3| \\ &\cdot |G_2 \cap G_4| \end{aligned} \right\} \leq \left\{ \begin{aligned} &|G_1| |G_1 \cap G_2| |G_3|^2 \\ &\cdot |G_4|^2 |G_1 \cap G_3 \cap G_4|^4, \\ &\cdot |G_2 \cap G_3 \cap G_4| \end{aligned} \right.$$

which holds for all finite group $|G|$ and subgroups $G_1, G_2, G_3,$ and G_4 . We call such an inequality a “non-Shannon-type” group inequality. Curiously, there has not been a proof of this inequality based on group theory alone (without going through the entropy function), which can shed light on the group-theoretic meaning of this inequality. Likewise, from any other non-Shannon-type entropy inequality, one can obtain the corresponding group inequality.

In the above, we have discussed how to obtain a group inequality from an entropy inequality. On the other hand, if a group inequality of the form (41) or (43) holds, then the corresponding entropy inequality of the form (38) or (42) also holds.

This one-to-one correspondence between entropy inequalities and group inequalities is intimately related to a combinatorial structure known as the *quasi-uniform array* [40]. This combinatorial structure, inspired by the fundamental notion of *strong typicality* in information theory, is exhibited by any finite group and its subgroups. We refer the reader to [41][4], Ch. 16] for the details.

5.4.4. Matrix Theory

Let X be a continuous random variable with probability density function (pdf) $f(x)$. The differential entropy of X is defined as

$$h(X) = - \int f(x) \log f(x) dx.$$

Likewise, the joint differential entropy of a random vector \mathbf{X} with joint pdf $f(\mathbf{x})$ is defined as

$$h(\mathbf{X}) = - \int f(\mathbf{x}) \log f(\mathbf{x}) d\mathbf{x}. \tag{44}$$

The integral in the above definitions are assumed to be taken over the support of the underlying pdf.

A linear differential entropy inequality

$$\sum_{\alpha \in \bar{N}} c_\alpha h(X_\alpha) \geq 0$$

is said to be balanced if for all $i \in [n]$, we have $\sum_{\alpha \in \bar{N}: i \in \alpha} c_\alpha = 0$. (The same can be defined for an entropy inequality.) It was proved in [42] that the above differential entropy inequality is valid if and only if it is balanced and its discrete analog is valid. For example,

$$h(X|Y) = h(X, Y) - h(Y) \geq 0$$

is not valid because it is not balanced. On the other hand,

$$I(X; Y) = h(X) + h(Y) - h(X, Y) \geq 0$$

is valid because it is balanced and its discrete analog

$$H(X) + H(Y) - H(X, Y) \geq 0$$

is valid. Thus if Γ_n^* can be determined, then in principle all valid differential entropy inequalities can be determined.

Any $n \times n$ symmetric positive semi-definite matrix $K = [k_{ij}]$ defines a Gaussian vector $\mathbf{X} = [X_1 X_2 \cdots X_n]$ with covariance matrix K . Substituting the corresponding Gaussian distribution into (44), we obtain

$$h(\mathbf{X}) = \frac{1}{2} \log[(2\pi e)^n |K|],$$

where $|\cdot|$ denotes the determinant of a matrix. For $\alpha \in \bar{N}$, let K_α be the submatrix of K at the intersection of the rows and the columns of K indexed by α , whose determinant $|K_\alpha|$ is called a *principal minor* of K . Note that K_α is the covariance matrix of the subvector $\mathbf{X}_\alpha = [X_i : i \in \alpha]$. Since \mathbf{X}_α is also Gaussian, it follows that

$$h(\mathbf{X}_\alpha) = \frac{1}{2} \log[(2\pi e)^{|\alpha|} |K_\alpha|]. \quad (45)$$

Now consider the independence bound for differential entropy,

$$h(X_1, X_2, \dots, X_n) \leq \sum_i h(X_i),$$

which is tight if and only if $X_i, i \in [n]$ are mutually independent. Substituting (45) into the above, we have

$$\frac{1}{2} \log[(2\pi e)^n |K|] \leq \sum_i \frac{1}{2} \log[(2\pi e) K_i],$$

or

$$\frac{n}{2} \log(2\pi e) + \frac{1}{2} \log |K| \leq \frac{n}{2} \log(2\pi e) + \frac{1}{2} \log \prod_i K_i.$$

Note that those terms involving $\frac{1}{2} \log(2\pi e)$ are cancelled out, because the independence bound is a valid differential entropy inequality and so it is balanced. After simplification, we obtain

$$|K| \leq \prod_i K_i,$$

namely *Hadamard's inequality*, which is tight if and only if $X_i, i \in [n]$ are mutually independent, or $k_{ij} = 0$ for all $i \neq j$.

This and similar techniques can be applied to obtain various inequalities on the principal minors of symmetric positive semi-definite matrices [8], Section 16.8]. These include a generalization of Hadamard's inequality due to Szász [43] and the Minkowski inequality [44].

For every valid differential entropy inequality, a corresponding inequality involving the principal minors of a symmetric positive semi-definite matrix can be obtained in this fashion. It turns out that all non-Shannon-type inequalities for discrete random variables discovered so far are balanced, and so they are also valid for differential entropy. For example, from ZY98 we can obtain

$$|K_1| |K_{12}| |K_3|^2 |K_4|^2 |K_{134}|^4 |K_{234}| \leq |K_{13}|^3 |K_{14}|^3 |K_{34}|^3 |K_{23}| |K_{24}|,$$

which can be called a “non-Shannon-type” inequality for 4×4 positive semi-definite matrix K . It was proved in [45] that for 3×3 positive semi-definite matrices, all inequalities involving the principal minors can be obtained through the Gaussian distribution as explained.

5.4.5. Kolmogorov Complexity

Kolmogorov complexity, also known as Kolmogorov-Chatin complexity, is a subfield of computer science. The Kolmogorov complexity of a sequence x , denoted by $K(x)$, is the length of the shortest description of the string with respect to a *universal description language*. Without getting into the details, such a universal description language can be based on a computer programming language. Likewise, the Kolmogorov complexity of a pair of sequences x and y is denoted by $K(x, y)$. We refer the reader to [46] for a comprehensive treatment of the subject.

Hammer *et al.* [47] established that all linear inequalities that are valid for Kolmogorov complexity are also valid for entropy, and vice versa. For example, the inequality

$$H(X_1) + H(X_2) \geq H(X_1, X_2)$$

for any X_1, X_2 corresponds to the inequality

$$K(x_1) + K(x_2) \geq K(x_1, x_2)$$

for any two sequences x_1 and x_2 . This establishes a one-to-one correspondence between entropy and Kolmogorov complexity. Due to this one-to-one correspondence, “non-Shannon-type” inequalities for Kolmogorov complexity can be obtained accordingly.

5.4.6. Quantum Mechanics

The von Neumann entropy ^[48] is a generalization of the classical entropy (Shannon entropy) to the field of quantum mechanics.¹⁰ For any quantum state described by a Hermitian positive semi-definite matrix ρ , the von Neumann entropy of ρ is defined as

$$S(\rho) = -\text{Tr}(\rho \log \rho).$$

Consider distinct quantum systems A and B . The joint system is described by a Hermitian positive semi-definite matrix ρ_{AB} . The individual systems are described by ρ_A and ρ_B which are obtained from ρ_{AB} by taking partial trace. Consider a fixed ρ_{AB} . We simply use $S(A)$ to denote the entropy of System A , i.e., $S(\rho_A)$. In the following, the same convention applies to other joint or individual systems. It is well known that

$$|S(A) - S(B)| \leq S(AB) \leq S(A) + S(B).$$

The second inequality above is called the *subadditivity* for the von Neumann entropy. The first inequality, called the *triangular inequality* (also known as the Araki-Lieb inequality ^[49]), is regarded as the quantum analog of the inequality

$$H(X) \leq H(X, Y) \tag{46}$$

for the Shannon entropy. It is important to note that although the Shannon entropy of a joint system is always not less than the Shannon entropy of an individual system as shown in (46), this may not be true in quantum systems. It is possible that $S(AB) = 0$ but $S(A) > 0$ and $S(B) > 0$, for example, when AB is a pure entangled state ^[50]. From this fact, we can see that the quantum world can be quite different from the classical world.

The *strong subadditivity* of the von Neumann entropy ^{[51][52]} plays the same role as the basic inequalities for the classical entropy. For distinct quantum systems A , B , and C , strong subadditivity can be represented by the following two equivalent forms:

$$\begin{aligned} S(A) + S(B) &\leq S(AC) + S(BC) \\ S(ABC) + S(B) &\leq S(AB) + S(BC). \end{aligned}$$

These inequalities can be used to show many other interesting inequalities involving conditional entropy and mutual information. Similar to classical information theory, quantum conditional entropy and quantum mutual information are defined as $S(A|B) = S(A, B) - S(B)$ and $S(A : B) = S(A) + S(B) - S(A, B)$, respectively. For distinct quantum systems A , B , C and D , we have ^[50]

i) *Conditioning reduces conditional entropy:*

$$S(A|B, C) \leq S(A|B).$$

ii) *Discarding quantum systems never increases mutual information:*

$$S(A : B) \leq S(A : B, C).$$

iii) *Subadditivity of conditional entropy* ^[53]:

$$\begin{aligned} S(A, B|C, D) &\leq S(A|C) + S(B|D) \\ S(A, B|C) &\leq S(A|C) + S(B|C) \\ S(A|B, C) &\leq S(A|B) + S(A|C). \end{aligned}$$

Following the discovery of non-Shannon-type inequalities for the classical entropy, it became natural to ask whether there exist constraints on the von Neumann entropy beyond strong subadditivity. It was proved a few years later that for a three-party system, there exist no such constraint ^[54]. Subsequently, a constrained inequality for the von Neumann entropy for a four-party system which is independent of strong subadditivity was discovered ^[55], and a family of countably infinitely many constrained inequalities that are independent of each other and strong subadditivity was proved ^[56].

6. Concluding Remarks

In this paper, we have developed a framework for universally quantified inequalities. With its root in information theory, this framework provides a geometrical interpretation that captures the very meaning of such inequalities. With this formality, we have revisited three celebrated inequalities in mathematics, namely the AM-GM inequality, Markov's inequality, and the Cauchy-Schwarz inequality, and clarified the related issues. To demonstrate the power of this formality, we have discussed its application to the study of entropy inequalities that have yielded very fruitful results in a number of subjected related to the Shannon entropy. Application of this formality to different branches of mathematics can identify situations in which new fundamental inequalities on quantities of interest may exist, and potentially lead to the discovery of such inequalities.

Footnotes

¹ The discussion in this section is built upon the concept of “achievability”, which will continue to play a central role in the rest of the paper. We note that *achievability* is a fundamental concept in information theory for formulating coding theorems. For example, C. E. Shannon’s celebrated source coding theorem [\[1\]](#) states that the coding rate for lossless data compression must be at least equal to the entropy rate of the information source. This means that any coding rate above the entropy rate of the information source is asymptotically achievable by some encoding–decoding schemes.

² For any $A \subset \mathbb{R}^2$, if we let

$$f(a, g) = \begin{cases} 1 & \text{if } (a, g) \in A \\ -1 & \text{if } (a, g) \notin A, \end{cases}$$

then $(a, g) \in A$ if and only if $f(a, g) \geq 0$. Thus an inequality of the form (3) can constrain (AM, GM) to any subset of \mathbb{R}^2 .

³ In the literature there have been works on sharpening (or refining) the Cauchy-Schwarz inequality, for example [\[57\]\[58\]](#). The bounds on $\langle \mathbf{u}, \mathbf{v} \rangle$ obtained in these works are tighter than the Cauchy-Schwarz inequality. However, these bounds depend on quantities other than $\langle \mathbf{u}, \mathbf{u} \rangle$ and $\langle \mathbf{v}, \mathbf{v} \rangle$, and so they are beyond the scope of the current work. Nevertheless, for such a bound, a geometrical framework similar to the one discussed in this section can be introduced, but it would be more complicated because more than three quantities are involved. Generally speaking, if more information about \mathbf{u} and \mathbf{v} is available, then tighter bounds on $\langle \mathbf{u}, \mathbf{v} \rangle$ can be obtained. In particular, if full information about \mathbf{u} and \mathbf{v} is available, i.e., both \mathbf{u} and \mathbf{v} are known, then $\langle \mathbf{u}, \mathbf{v} \rangle$ can be determined exactly.

⁴ A general discussion on the Shannon entropy and related inequalities can be found in a blog by Terence Tao [\[59\]](#).

⁵ Equivalently, $\mathbf{h} \in \mathcal{H}_n$ is entropic if it is achievable by some collection of n random variables.

⁶ For a convex polytope $P \subseteq \mathcal{H}_n$, a face is any set of the form $F = P \cap \{\mathbf{h} \in \mathcal{H}_n : \mathbf{b}^\top \mathbf{h} = c\}$, where $\mathbf{b}^\top \mathbf{h} \leq c$ for all $\mathbf{h} \in P$.

⁷ In network communication, multicast refers to transmitting a message to a specified subset of destination nodes in the network.

⁸ The capacity region contains all the achievable information rate tuples of the network coding problem.

⁹ See for example [44], Theorem 16.27] for a proof.

¹⁰ We refer the reader to [50] for a comprehensive treatment of quantum information theory.

References

1. ^aC. E. Shannon, "A Mathematical Theory of Communication," *Bell Sys. Tech. Journal*, 27: 379-423, 623-656, 1948.
2. ^aR. W. Yeung, "A framework for linear information inequalities," *IEEE Trans. Info. Theory*, IT-43: 1924-1934, 1997.
3. [^]AM-GM inequality, https://en.wikipedia.org/wiki/AM-GM_inequality
4. ^a^b^c^d^eR. W. Yeung, *Information Theory and Network Coding*, Springer 2008.
5. ^aR. W. Yeung, "Facets of entropy," *Communications in Information and Systems*, vol. 15, no. 1, 87-117, 2015.
6. ^a^bT. Chan, "Recent progresses in characterizing information inequalities," *Entropy*, 13: 379-401, 2011.
7. [^]R. G. Gallager, *Information Theory and Reliable Communication*, Wiley, New York, 1968.
8. ^a^bT. M. Cover and J. A. Thomas, *Elements of Information Theory*, Wiley, 1991, 2nd ed., Wiley-Interscience, 2006.
9. ^a^bN. Pippenger, "What are the laws of information theory?" 1986 Special Problems on Communication and Computation Conference, Palo Alto, CA, Sept. 3-5, 1986.
10. [^]S. Fujishige, "Polymatroidal dependence structure of a set of random variables," *Info. Contr.*, 39: 55-72, 1978.
11. [^]R. W. Yeung and Y.-O. Yan, *Information-Theoretic Inequality Prover (ITIP)*, 1996, <http://user-www.ie.cuhk.edu.hk/~ITIP/>
12. [^]S.-W. Ho, L. Ling, C. W. Tan, and R. W. Yeung, *AITIP*, 2020. Available: <https://github.com/convexsoft/AITIP>
13. [^]C. T. Li, 2021. Available: <https://github.com/cheuktingli/psitip>
14. [^]A. El Gamal and Y.-H. Kim, *Network Information Theory*, Cambridge University Press 2011.
15. [^]IEEE Information Theory Society: Software. <https://www.itsoc.org/resources/software>
16. [^]L. Guo, R. W. Yeung, and X.-S. Gao, "Proving information inequalities and identities with symbolic computation," *IEEE Trans. Info. Theory*, IT-69: 4799-4811, 2023.
17. [^]L. Guo, R. W. Yeung, and X.-S. Gao, "Proving information inequalities by Gaussian elimination," to appear in *IEEE Trans. Info. Theory*.

18. [△]R. W. Yeung, and C. T. Li, "Machine-proving of entropy inequalities," *IEEE BITS the Information Theory Magazine*, 1: 12-22, 2021.
19. [△]R. W. Yeung, "A new outlook on Shannon's information measures," *IEEE Trans. Info. Theory*, IT-37: 466-474, 1991.
20. [△]a. [△]b. [△]c. [△]F. Matúš, "Two constructions on limits of entropy functions," *IEEE Trans. Info. Theory*, IT-53: 320-330, 2007.
21. [△]Z. Zhang and R. W. Yeung, "A non-Shannon-type conditional inequality of information quantities," *IEEE Trans. Info. Theory*, IT-43: 1982-1986, 1997.
22. [△]Z. Zhang and R. W. Yeung, "On characterization of entropy function via information inequalities," *IEEE Trans. Info. Theory*, IT-44: 1440-1452, 1998.
23. [△]K. Makarychev, Y. Makarychev, A. Romashchenko, and N. Vereshchagin, "A new class of non-Shannon-type inequalities for entropies," *Comm. Info. and Syst.*, 2: 147-166, 2002.
24. [△]Z. Zhang, "On a new non-Shannon-type information inequality," *Comm. Info. and Syst.*, 3: 47-60, 2003.
25. [△]W. Xu, J. Wang, and J. Sun, "A projection method for derivation of non-Shannon-type information inequalities," *IEEE International Symposium on Information Theory*, Toronto, Jul 6-11, 2008.
26. [△]R. Dougherty, C. Freiling, and K. Zeger, "Non-Shannon information inequalities in four random variables," [Online]: <http://arxiv.org/abs/1104.3602> (Apr. 2011).
27. [△]L. Csirmaz, "Book inequalities," *IEEE Trans. Info. Theory*, IT-60: 6811-6818, 2014.
28. [△]F. Matúš, "Piecewise linear conditional information inequality," *IEEE Trans. Info. Theory*, IT-52: 236-238, 2006.
29. [△]Q. Chen and R. W. Yeung, "Characterizing the entropy function region via extreme rays," *2012 IEEE Information Theory Workshop*, Lausanne, Switzerland, Sept 3-7, 2012.
30. [△]Q. Chen, M. Cheng, and B. Bai, "Matroidal entropy functions: A quartet of theories of information, matroid, design and coding," *Entropy*, 23: 1-11, 2021.
31. [△]S. Liu and Q. Chen, "Entropy functions on two-dimensional faces of polymatroidal region of degree four," *2023 IEEE International Symposium on Information Theory*, Taipei, Taiwan, Jun 25-30, 2023.
32. [△]Q. Chen, M. Cheng, and B. Bai, "Matroidal entropy functions: Constructions, characterizations and representations," to appear in *IEEE Trans. Info. Theory*.
33. [△]a. [△]b. [△]R. W. Yeung and Z. Zhang, "Distributed source coding for satellite communications," *IEEE Trans. Info. Theory*, IT-45: 1111-1120, 1999.

34. ^aR. Ahlswede, N. Cai, S.-Y. R. Li, and R. W. Yeung, "Network information flow," *IEEE Trans. Info. Theory*, IT-46: 1204-1216, 2000.
35. ^aX. Yan, R. W. Yeung, and Z. Zhang, "An implicit characterization of the achievable rate region for acyclic multisource multisink network coding," *IEEE Trans. Info. Theory*, IT-58: 5625-5639, 2012.
36. ^aL. Song, R. W. Yeung and N. Cai, "A separation theorem for single-source network coding," *IEEE Trans. Info. Theory*, IT-52: 1861-1871, 2006.
37. ^aF. Matúš, "Conditional independences among four random variables III: Final conclusion," *Combinatorics, Probability and Computing*, 8: 269-276, 1999.
38. ^aF. Matúš and M. Studený, "Conditional independences among four random variables I," *Combinatorics, Probability and Computing*, 4: 269-278, 1995.
39. ^aF. Matúš, "Conditional independences among four random variables II," *Combinatorics, Probability and Computing*, 4: 407-417, 1995.
40. ^aT. H. Chan, "A combinatorial approach to information inequalities," *Comm. Info. and Syst.*, 1: 241-253, 2001.
41. ^aT. H. Chan and R. W. Yeung, "On a relation between information inequalities and group theory," *IEEE Trans. Info. Theory*, IT-48: 1992-1995, 2002.
42. ^aT. H. Chan, "Balanced information inequalities," *IEEE Trans. Info. Theory*, IT-49: 3261-3267, 2003.
43. ^aL. Mirsky, "On a generalization of Hardamad's determinantal inequality due to Szász," *Arch. Math.*, VIII: 274-275, 1957.
44. ^aH. Minkowski, "Diskontinuitätsbereich für arithmetische Äquivalenz," *Journal für Math.*, 129: 220-274, 190.
45. ^aT. Chan, D. Guo, and R. Yeung, "Entropy functions and determinant inequalities," 2012 *IEEE International Sym. on Info. Theory*, Cambridge, MA, USA, Jul 1-6, 2012.
46. ^aM. Li and P. Vitányi, *An Introduction to Kolmogorov Complexity and Its Applications*, 3rd ed., Springer, New York, 2008.
47. ^aD. Hammer, A. Romashchenko, A. Shen, and N. Vereshchagin, "Inequalities for Shannon Entropy and Kolmogorov Complexity," *J. Comp. and Syst. Sci.*, 60: 442-464, 2000.
48. ^aJ. von Neumann, *Mathematische Grundlagen der Quantenmechanik*, Springer, Berlin, 1932.
49. ^aH. Araki and E. H. Lieb. "Entropy inequalities". *Comm. Math. Phys.*, 18:160-170, 1970.
50. ^aM. A. Nielsen and I. L. Chuang, *Quantum Computation and Quantum Information*, Cambridge University Press, 2000.

51. [△]E. H. Lieb and M. B. Ruskai, "A fundamental property of quantum-mechanical entropy," *Phys. Rev. Lett.*, 30 (10): 434–436, 1973.
52. [△]E. H. Lieb and M. B. Ruskai, "Proof of the strong subadditivity of quantum mechanical entropy," *J. Math. Phys.*, 14: 1938–1941, 1973.
53. [△]M. A. Nielsen. *Quantum Information Theory*. Ph.D. thesis, University of New Mexico, 1998.
54. [△]N. Pippenger, "The inequalities of quantum information theory," *IEEE Trans. Info. Theory*, IT-49, 773-789, 2003.
55. [△]N. Linden and A. Winter, "A new inequality for the von Neumann entropy," *Comm. Math. Phys.*, 259: 129-138, 2005.
56. [△]J. Cadney, N. Linden and A. Winter, "Infinitely many constrained inequalities for the von Neumann entropy," *IEEE Trans. Info. Theory*, IT-58: 3657-3663, 2012.
57. [△]J. W. Hovenier, "Sharpening Cauchy's inequality," *Journal of Mathematical Analysis and Applications*, 186: 156-160, 1993.
58. [△]P. R. Mercer, "A refined Cauchy-Schwarz inequality," *International Journal of Mathematical Education in Science and Technology*, 38: 839-843, 2009.
59. [△]Terence Tao's blog: Special cases of Shannon entropy. <https://terrytao.wordpress.com/2017/03/01/special-cases-of-shannon-entropy/#comment-479252>

Declarations

Funding: No specific funding was received for this work.

Potential competing interests: No potential competing interests to declare.