Communications in Information and Systems
Volume 15, Number 1, 87–117, 2015

Facets of entropy

Raymond W. Yeung

Constraints on the entropy function are of fundamental importance in information theory. For a long time, the polymatroidal axioms, or equivalently the nonnegativity of the Shannon information measures, are the only known constraints. Inequalities that are implied by nonnegativity of the Shannon information measures are categorically referred to as Shannon-type inequalities. If the number of random variables is fixed, a Shannon-type inequality can in principle be verified by a software package known as ITIP. A non-Shannon-type inequality is a constraint on the entropy function which is not implied by the nonnegativity of the Shannon information measures. In the late 1990's, the discovery of a few such inequalities revealed that Shannon-type inequalities alone do not constitute a complete set of constraints on the entropy function. In the past decade or so, connections between the entropy function and a number of subjects in information sciences, mathematics, and physics have been established. These subjects include probability theory, network coding, combinatorics, group theory, Kolmogorov complexity, matrix theory, and quantum mechanics. This expository work is an attempt to present a picture for the many facets of the entropy function.

Keywords and Phrases: Entropy, polymatroid, non-Shannon-type inequalities, positive definite matrix, quasi-uniform array, Kolmogorov complexity, conditional independence, network coding, quantum information theory.

This work was partially supported by a grant from the University Grants Committee of the Hong Kong Special Administrative Region, China (Project No. AoE/E-02/08), and also by the Shenzhen Key Laboratory of Network Coding Key Technology and Application (ZSDY20120619151314964) and the Peacock Scheme (KQCX2013062816400 8004), Shenzhen, China.

The work is based on the author's plenary talk with the same title at the 2009 IEEE International Symposium on Information Theory, Seoul, Korea, Jun 28 – Jul 3, 2009.

1. Preliminaries

Let $[n] = \{1, \ldots, n\}$, $\mathsf{N} = 2^{[n]}$, and $\overline{\mathsf{N}} = \mathsf{N} \backslash \{\emptyset\}$. Let $\Theta = \{X_i, i \in [n]\}$ be a collection of n discrete random variables. We will not discuss continuous random variables until Section 3.6, so unless otherwise specified, a random variable is assumed to be discrete. Let p_X denote the probability distribution of a random variable X. The entropy (Shannon entropy) [56] of X is defined by

$$H(X) = -\sum_x p_X(x) \log p_X(x).$$

The base of the logarithm is taken to be some convenient positive real number. When it is equal to 2, the unit of entropy is the *bit.* Likewise, the joint entropy of two random variables X and Y is defined by

$$H(X,Y) = -\sum_{x,y} p_{XY}(x,y) \log p_{XY}(x,y).$$

This definition is readily extendible to any finite number of random variables. All summations are assumed to be taken over the support of the underlying distribution. For example, for $H(X,Y)$ above, the summation is taken over all x and y such that $p_{XY}(x,y) > 0$.

Note that the quantity $H(X)$ is defined upon the distribution p_X and does not depend on the actually values taken by X. Therefore, we also write $H(p_X)$ for $H(X)$, $H(p_{XY})$ for $H(X,Y)$, etc.

In information theory, entropy is the measure of the uncertainty contained in a discrete random variable, justified by fundamental coding theorems. For comprehensive treatments of information theory, we refer the reader to [16, 24, 65].

For n random variables, there are $2^n - 1$ joint entropies. For example, for $n = 3$, the 7 joint entropies are

$$H(X_1), H(X_2), H(X_3), H(X_1,X_2), H(X_2,X_3), H(X_1,X_3), H(X_1,X_2,X_3).$$

For $\alpha \in \mathsf{N}$, 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 $H_\Theta : \mathsf{N} \to \Re$ by

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

with $H_\Theta(\emptyset) = 0$ because $X_\emptyset$ is a constant. H_Θ is called the *entropy function* of Θ.[1]

In information theory, in addition to entropy, the following information measures are defined:

Conditional Entropy

$$H(X|Y) = H(X,Y) - H(Y)$$

Mutual Information

$$I(X;Y) = H(X) + H(Y) - H(X,Y)$$

Conditional Mutual Information

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

Together with entropy, these are called the *Shannon information measures.* Note that all the Shannon information measures are linear combinations of entropies.

An information expression refers to a function of the Shannon information measures involving a finite number of random variables. Thus an information expression can be written as a function of entropies, called the *canonical form* of the information expression. The uniqueness of the canonical form of a linear information expression was first proved by Han [27] and independently by Csiszár and Körner [17]. The uniqueness of the canonical form of more general information expressions was proved by Yeung [63]. Therefore, to study constraints on the Shannon information measures, it suffices to study constraints on the entropy function.

2. Shannon-type and non-Shannon-type inequalities

Fujishige [23] showed that for any Θ, H_Θ satisfies the following properties, known as the *polymatroidal axioms* : For any $\alpha, \beta \in \mathsf{N}$,

i) $H_\Theta(\phi) = 0$;

ii) $H_\Theta(\alpha) \leq H_\Theta(\beta)$ if $\alpha \subset \beta$;

[1] Motivated by the consideration of the capacity of networks, Hassibi and Shadbakht [29] introduced the *normalized* entropy function and studied its properties. In their definition, $X_i, i \in [n]$ are assumed to have the same alphabet size N, and $H_\Theta(\alpha) = (\log N)^{-1} H(X_\alpha)$.

iii) $H_\Theta(\alpha) + H_\Theta(\beta) \geq H_\Theta(\alpha \cap \beta) + H_\Theta(\alpha \cup \beta)$.

On the other hand, it is well known that all Shannon information measures are nonnegative, i.e.,

$$\begin{aligned}\text{entropy} &\geq 0\\ \text{conditional entropy} &\geq 0\\ \text{mutual information} &\geq 0\\ \text{conditional mutual information} &\geq 0.\end{aligned}$$

These inequalities are referred to as the *basic inequalities* of information theory. Note that the nonnegativity of conditional mutual information implies all the other forms of basic inequalities, and is therefore the most general form of basic inequalities. It can be shown that the polymatroidal axioms on the entropy function are equivalent to the basic inequalities [64, App. 13.A].

When we say that the entropy function satisfies the polymatroidal axioms, it means that for any joint distribution defined for $X_1, X_2, \ldots, X_n$, the corresponding $2^n - 1$ joint entropies satisfy these axioms. The same interpretation applies when we say that a constraint on the entropy function is valid.

Constraints on the entropy function govern the "impossibilities" in information theory. The proofs of most converse coding theorems rely on such constraints. For a long time, the polymatroidal axioms were the only known constraints on the entropy function. In the 1980's, Pippenger [51][2] asked whether these exist constraints on the entropy function other than the polymatroidal axioms. He called constraints on the entropy function the *laws of information theory.* If there are additional constraints on the entropy function, then perhaps new converse coding theorems can be proved.

In the 1990's, Yeung [63] studied constraints on the entropy function and introduced the following geometrical formulation of the problem. First, the number of random variables n is fixed to be some positive integer. Compared with [51], this makes the setting of the problem finite dimensional instead of infinite dimensional, and hence more manageable. Let $\mathcal{H}_n \overset{\triangle}{=} \Re^{2^n-1}$, where the coordinates of $\mathcal{H}_n$ are labeled by $h_\alpha, \alpha \in \overline{\mathsf{N}}$. We call $\mathcal{H}_n$ the *entropy space* for n random variables. Then for each collection Θ of n random variables, H_Θ can be represented by a 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 $\alpha \in \overline{\mathsf{N}}$. On the other hand, a vector $\mathbf{h} \in \mathcal{H}_n$ is called *entropic* if it is equal to the entropy

[2] The author would like to thank Prof. Nick Pippenger for pointing out his work.

vector of some collection Θ of n random variables. Define the following region in $\mathcal{H}_n$:

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

The region Γ_n^*, or simply Γ^* when n is not specified, is referred to as the region of entropy functions. If Γ_n^* can be determined, then in principle all valid entropy inequalities can be determined.

Consider an entropy inequality of the form $f(\mathbf{h}) \geq 0$.[3] For example, the inequality

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

corresponds to $f(\mathbf{h}) \geq 0$ with $f(\mathbf{h}) = h_1 + h_2 - h_{12}$. The above setup enables constraints on the entropy function to be interpreted geometrically. Specifically, an entropy inequality $f(\mathbf{h}) \geq 0$ is valid if and only if

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

In fact, $f(\mathbf{h}) \geq 0$ is valid if and only if

$$\overline{\Gamma}_n^* \subset \{\mathbf{h} \in \mathcal{H}_n : f(\mathbf{h}) \geq 0\}$$

because $\{\mathbf{h} \in \mathcal{H}_n : f(\mathbf{h}) \geq 0\}$ is closed. Figure 1 (a) and (b) illustrates the two possible scenarios for $f(\mathbf{h}) \geq 0$.

In information theory, we very often deal with information inequalities with certain constraints on the joint distribution of the random variables involved. These are called constrained information inequalities, and the constraints on the joint distribution can usually be expressed as linear constraints on the entropies. For example, $X_1 \to X_2 \to X_3 \to X_4$ forms a Markov chain if and only if $I(X_1; X_3|X_2) = 0$ and $I(X_1, X_2; X_4|X_3) = 0$. Under this Markov constraint, $I(X_2; X_3) \geq I(X_1; X_4)$, called the *data processing inequality*, is well known.

We now define another region Γ_n in $\mathcal{H}_n$ that corresponds to the basic inequalities (for n random variables):

$$\Gamma_n = \{\mathbf{h} \in \mathcal{H}_n : \mathbf{h} \text{ satisfies the basic inequalities}\}.$$

(The region Γ_n is written as Γ when n is not specified.) Note that Γ_n is a polytope in the positive orthant of $\mathcal{H}_n$ (and so it is computable), and $\Gamma_n^* \subset \Gamma_n$ because the basic inequalities are satisfied by any $X_1, X_2, \ldots, X_n$.

[3] We consider only non-strict inequalities because these are the inequalities usually used in information theory.

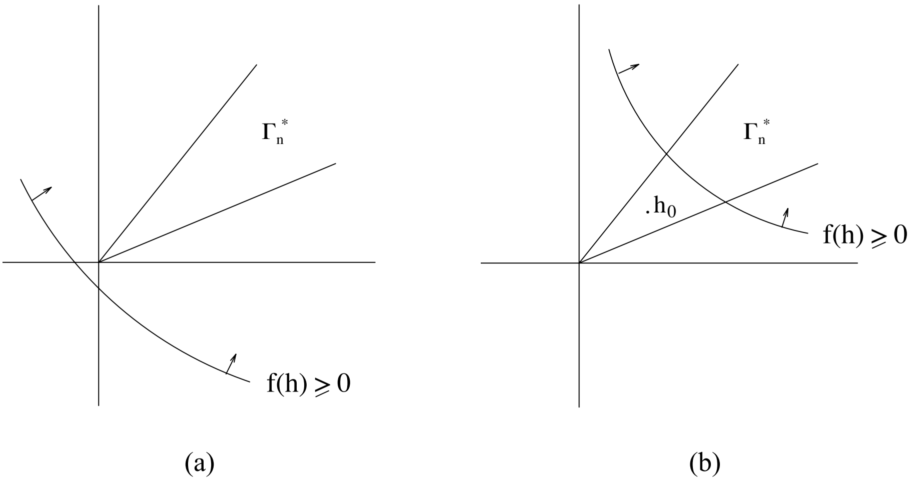

Figure 1: (a) Γ_n^* is contained in $\{\mathbf{h} \in \mathcal{H}_n : f(\mathbf{h}) \geq 0\}$. (b) Γ_n^* is not contained in $\{\mathbf{h} \in \mathcal{H}_n : f(\mathbf{h}) \geq 0\}$. In this case, there exists an entropy vector $\mathbf{h}_0$ that does not satisfy $f(\mathbf{h}) \geq 0$.

An entropy inequality $f(\mathbf{h}) \geq 0$ is called a *Shannon-type inequality* if it is implied by the basic inequalities, or

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

Constrained Shannon-type inequalities, namely those constrained inequalities that are implied by the basic inequalities, can also be formulated in terms of Γ_n [63].

This formulation of Shannon-type inequalities enables machine proving of such inequalities (both unconstrained and constrained), namely that a Shannon-type inequality can be verified by solving a linear program. See [63] for a detailed discussion. ITIP, a software package for this purpose that runs on MATLAB, was developed by Yeung and Yan [68]. A platform-independent version of ITIP that runs on C, called Xitip, was developed by Pulikkoonattu et al. [53]. Another software package for the same purpose that is axiom based was developed by Chung [15].

With ITIP, there is now a way to determine whether an entropy inequality is Shannon-type or not. Specifically, if an inequality can be verified by ITIP, then it is a Shannon-type inequality, otherwise it is not. Thus we are in a position to discuss whether there exist entropy inequalities beyond

Shannon-type inequalities. If so, these inequalities would be called non-Shannon-type inequalities.

Let $\overline{\Gamma}_n^*$ denote the closure of Γ_n^*. Zhang and Yeung [71] proved the following fundamental properties of the region Γ_n^*:

i) $\Gamma_2^* = \Gamma_2$;

ii) $\Gamma_3^* \neq \Gamma_3$, but $\overline{\Gamma}_3^* = \Gamma_3$;[4]

iii) For $n \geq 3$, Γ_n^* is neither closed nor convex, but $\overline{\Gamma}_n^*$ is a convex cone.

Therefore, unconstrained non-Shannon-type inequalities can exist only for 4 or more random variables. In the same work, the following constrained non-Shannon-type inequality for 4 random variables was proved.

Theorem 1 (ZY97). *For any four random variables X_1, X_2, X_3, and X_4, if $I(X_1; X_2) = I(X_1; X_2|X_3) = 0$, then*

$$I(X_3; X_4) \leq I(X_3; X_4|X_1) + I(X_3; X_4|X_2).$$

The inequality ZY97 implies the existence of a non-entropic region on the boundary of Γ_4. However, this alone is not sufficient to establish that $\overline{\Gamma}_4^*$ is strictly smaller than Γ_4. Shortly afterwards, Zhang and Yeung [72] proved the following unconstrained non-Shannon-type inequality for 4 random variables, showing that indeed $\overline{\Gamma}_4^* \neq \Gamma_4$.

Theorem 2 (ZY98). *For any four random variables X_1, X_2, X_3, and X_4,*

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

The inequality ZY98 cut through the gap between $\overline{\Gamma}_4^*$ and Γ_4. This is illustrated in Figure 2.

This inequality has been further generalized by Makarychev et al. [40], Zhang [70], and Matúš [45]. In particular, Matúš showed that $\overline{\Gamma}_n^*$ is not a polyhedral cone, and hence there exist infinitely many linear non-Shannon-type inequalities. On the other hand, by modifying ITIP, Dougherty et al. [19] have discovered a number of non-Shannon-type inequalities by a search on a supercomputer.

[4] Previously, Han [28] proved that Γ_3 is the smallest cone that contains Γ_3^*. The result $\overline{\Gamma}_3^* = \Gamma_3$ was also proved by Golić [25], and is also a consequence of the theorem in Matúš [41].

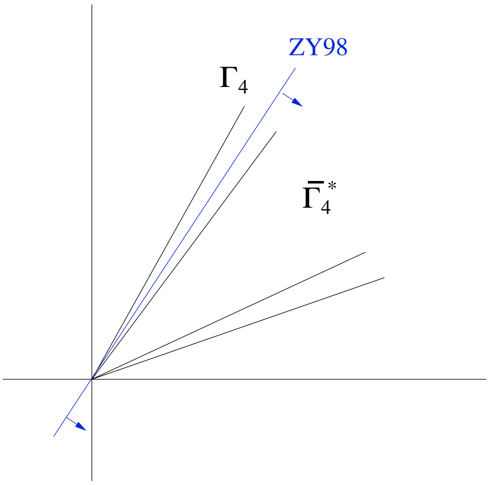

Figure 2: An illustration of non-Shannon-type inequality ZY98.

3. Connections with information sciences, mathematics, and physics

The study of constraints on the entropy function was originally motivated by information theory, but subsequent to the discovery of the first non-Shannon-type inequalities, fundamental connections have been made between information theory and various branches of information sciences, mathematics, and physics. These connections reveal "non-Shannon-type" inequalities for finite groups, Kolmogorov complexity, and positive definite matrices. Inspired by the existence of non-Shannon-type inequalities for the Shannon entropy, new inequalities have been discovered for the von Neumann entropy. In this section, we give a guided tour for each of these connections. We also refer the reader to Chan [11] for a more in-depth discussion.

3.1. Combinatorics

Consider a finite alphabet $\mathcal{X}$. For a sequence $\mathbf{x} \in \mathcal{X}^n$, let $N(x;\mathbf{x})$ be the number of occurrences of x in $\mathbf{x}$, and let $q(x) = n^{-1}N(x;\mathbf{x})$. The distribution $q_{\mathbf{x}} = \{q(x)\}$ is called the *empirical distribution* of $\mathbf{x}$.

Central in information theory is the notion of *typical sequences* with respect to a probability distribution defined on some alphabet. Consider any probability distribution p_X on $\mathcal{X}$. Roughly speaking, we say that a sequence

$\mathbf{x} \in \mathcal{X}^n$ is typical with respect to p_X if its empirical distribution manifests in some way the distribution p_X. There are different ways to measure the typicality of a sequence. Here we focus on the notion of *strong typicality* [4, 17, 59], and we adopt the definitions in [65].[5] Detailed discussions of the related fundamental results can be found therein.

Definition 1. The strongly typical set $T^n_{[X]\delta}$ with respect to p_X is the set of sequences $\mathbf{x} \in \mathcal{X}^n$ such that $N(x; \mathbf{x}) = 0$ for all x with $p(x) = 0$, and

$$\| p_X - q_{\mathbf{x}} \| \leq \delta,$$

where $\| \cdot \|$ denotes the L^1-norm, and δ is an arbitrarily small positive real number. The sequences in $T^n_{[X]\delta}$ are called strongly δ-typical sequences.

This definition can readily be extended to the bivariate case. Here we consider a joint alphabet $\mathcal{X} \times \mathcal{Y}$, a joint probability distribution p_{XY} on $\mathcal{X} \times \mathcal{Y}$, and sequences $(\mathbf{x}, \mathbf{y}) \in \mathcal{X}^n \times \mathcal{Y}^n$. The notations we use for the single-variate case are extended naturally. It suffices to say that a pair of sequences $(\mathbf{x}, \mathbf{y}) \in \mathcal{X}^n \times \mathcal{Y}^n$ is jointly δ-typical with respect to p_{XY} if $N(x, y; \mathbf{x}, \mathbf{y}) = 0$ for all (x, y) such that $p(x, y) = 0$ and $\| p_{XY} - q_{XY} \| \leq \delta$, and the strongly jointly typical set is denoted by $T^n_{[XY]\delta}$. Further extension to the multivariate case is straightforward.

For convenience, we write $H(X)$ for $H(p_X)$, $H(Y)$ for $H(p_Y)$, and $H(X, Y)$ for $H(p_{XY})$. By the *strong asymptotic equipartition property* (strong AEP), for sufficiently large n, $|T^n_{[X]\delta}| \approx 2^{nH(X)}$, $|T^n_{[Y]\delta}| \approx 2^{nH(Y)}$, and $|T^n_{[XY]\delta}| \approx 2^{nH(X,Y)}$. By the consistency property of strong typicality, if $(\mathbf{x}, \mathbf{y}) \in T^n_{[XY]\delta}$, then $\mathbf{x} \in T^n_{[X]\delta}$ and $\mathbf{y} \in T^n_{[Y]\delta}$.

Then the following becomes evident. Since there are $\approx 2^{nH(X,Y)}$ typical $(\mathbf{x}, \mathbf{y})$ pairs and $\approx 2^{nH(X)}$ typical $\mathbf{x}$, for a typical $\mathbf{x}$, the number of $\mathbf{y}$ such that $(\mathbf{x}, \mathbf{y})$ is jointly typical is

$$\approx \frac{2^{nH(X,Y)}}{2^{nH(X)}} = 2^{nH(Y|X)}$$

on the average. The conditional strong AEP further asserts that this not only is true on the average, but in fact is true for every typical $\mathbf{x}$ as long as

[5] The discussion here is based on strong typicality which applies only to random variables with finite alphabets. Recently, Ho and Yeung [32] introduced the notion of *unified typicality*, with which the same discussion can be applied to random variables with countable alphabets.

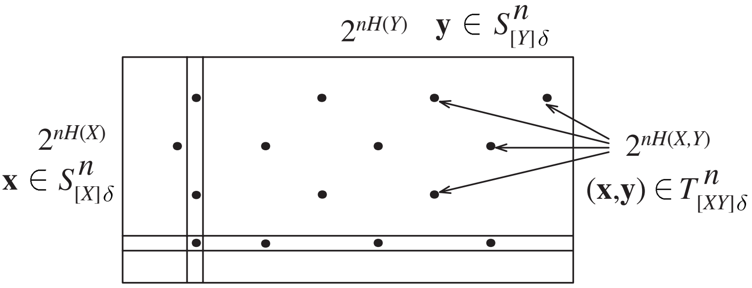
$2^{nH(Y)}$ $\mathbf{y} \in S^n_{[Y]\delta}$
$2^{nH(X)}$
$\mathbf{x} \in S^n_{[X]\delta}$
$2^{nH(X,Y)}$
$(\mathbf{x},\mathbf{y}) \in T^n_{[XY]\delta}$

Figure 3: A two-dimensional strong joint typicality array.

there exists one $\mathbf{y}$ such that $(\mathbf{x}, \mathbf{y})$ is jointly typical. Let $S^n_{[X]\delta}$ be the set of all such typical $\mathbf{x}$ sequences. The set $S^n_{[Y]\delta}$ is defined likewise.

We have established a rich set of structural properties for strong typicality with respect to a bivariate distribution p_{XY}, which is summarized in the two-dimensional *strong joint typicality array* in Figure 3. In this array, the rows and the columns are the typical sequences $\mathbf{x} \in S^n_{[X]\delta}$ and $\mathbf{y} \in S^n_{[Y]\delta}$, respectively. The total number of rows and columns are $\approx 2^{nH(X)}$ and $\approx 2^{nH(Y)}$, respectively. An entry indexed by $(\mathbf{x}, \mathbf{y})$ receives a dot if $(\mathbf{x}, \mathbf{y})$ is strongly jointly typical. The total number of dots is $\approx 2^{nH(X,Y)}$. The number of dots in each row is $\approx 2^{nH(Y|X)}$, while the number of dots in each column is $\approx 2^{nH(X|Y)}$.

From the strong typicality array, we see that the number of dots in the array is at most equal to the number of entries in the array, i.e.,

$$2^{nH(X,Y)} \leq 2^{nH(X)} 2^{nH(Y)}.$$

Upon taking the logarithm in the base 2 and dividing by n, we obtain

$$H(X,Y) \leq H(X) + H(Y),$$

or

$$I(X;Y) \geq 0.$$

Thus the basic inequality $I(X;Y) \geq 0$ is about the potentially unfilled entries in the two-dimensional strong typicality array.

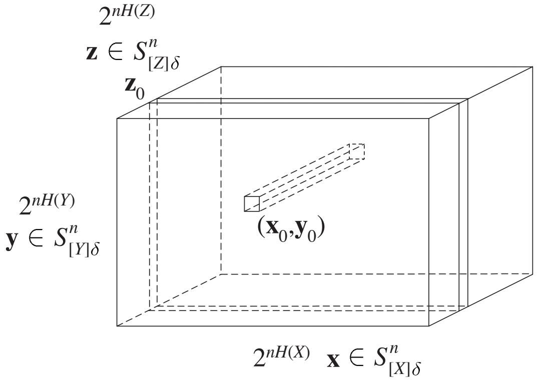
$2^{nH(Z)}$
$\mathbf{z} \in S^n_{[Z]\delta}$
$\mathbf{z}_0$
$2^{nH(Y)}$
$\mathbf{y} \in S^n_{[Y]\delta}$
$(\mathbf{x}_0,\mathbf{y}_0)$
$2^{nH(X)}$ $\mathbf{x} \in S^n_{[X]\delta}$

Figure 4: A three-dimensional strong joint typicality array.

We say that the strong joint typicality array in Figure 3 exhibits an *asymptotic quasi-uniform* structure. By a two-dimensional asymptotic quasi-uniform structure, we mean that in the array all the columns have approximately the same number of dots, and all the rows have approximately the same number of dots.

The strong joint typicality array for a multivariate distribution continues to exhibit an asymptotic quasi-uniform structure. Figure 4 shows a three-dimensional strong joint typicality array with respect to a distribution p_{XYZ}. As before, an entry $(\mathbf{x}, \mathbf{y}, \mathbf{z})$ receives a dot if $(\mathbf{x}, \mathbf{y}, \mathbf{z})$ is strongly jointly typical. This is not shown in the figure otherwise it will be very confusing. The total number of dots in the whole array is $\approx 2^{nH(X,Y,Z)}$. These dots are distributed in the array such that all the planes parallel to each other have approximately the same number of dots, and all the cylinders parallel to each other have approximately the same number of dots. More specifically, the total number of dots on the plane for any fixed $\mathbf{z}_0 \in S^n_{[Z]\delta}$ (as shown) is $\approx 2^{nH(X,Y|Z)}$, and the total number of dots in the cylinder for any fixed $(\mathbf{x}_0, \mathbf{y}_0)$ pair in $S^n_{[XY]\delta}$ (as shown) is $\approx 2^{nH(Z|X,Y)}$, so on and so forth. By investigating this array, it is not difficult to show that

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

which is the most general form of a basic inequality.

The discussion above gives a combinatorial interpretation of the basic inequalities. It is natural to ask whether all constraints on the entropy function, including non-Shannon-type inequalities, can be obtained by using this approach. Ideas along this line were further developed by Chan [9], where a *quasi-uniform array* was formally defined (to be elaborated in Section 3.2) and it was showed that all constraints on the entropy function can indeed be obtained through such arrays, and vice versa. This establishes a one-to-one correspondence between entropy and the combinatorial structure of a quasi-uniform array.

3.2. Group theory

Let G be a finite group with operation "$\circ$", and $G_1, G_2, \ldots, G_n$ be subgroups of G. Then for any $\alpha \in \overline{\mathsf{N}}$, $G_\alpha = \cap_{i\in\alpha} G_i$ is also a subgroup. For a group element a and a subgroup S, let aS denotes the left coset $a \circ S = \{a \circ s : s \in S\}$. In this section, we explain a one-to-one correspondence between entropy and finite groups established by Chan and Yeung [14]. The following lemma is instrumental.

Lemma 1. *Let G_i be subgroups of a group G and a_i be elements of G, $i \in \alpha$. Then*

$$\left|\bigcap_{i\in\alpha} a_i G_i\right| = \begin{cases} |G_\alpha| & \textit{if } \bigcap_{i\in\alpha} a_i G_i \neq \emptyset \\ 0 & \textit{otherwise} \end{cases}.$$

The meaning of this lemma can be explained by a simple example. The relation between a finite group G and subgroups G_1 and G_2 is illustrated by the *membership table* in Figure 5. In this table, an element of G is represented by a dot. The first column represents the subgroup G_1, with the dots in the first column being the elements in G_1. The other columns represent the left cosets of G_1. By Lagrange's theorem, all cosets of G_1 have the same order, and so all the columns have the same number of dots. Similarly, the first row represents the subgroup G_2 and the other rows represent the left cosets of G_2. Again, all the rows have the same number of dots.

The upper left entry in the table represents the subgroup $G_1 \cap G_2$. There are $|G_1 \cap G_2|$ dots in this entry, with one of them representing the identity element. Any other entry represents the intersection between a left coset of G_1 and a left coset of G_2, and by Lemma 1, the number of dots in each of these entries is either equal to $|G_1 \cap G_2|$ or zero.

We have already seen a similar structure in Figure 3 for the two-dimensional strong joint typicality array. In that array, when n is large, all the columns have approximately the same number of dots and all the rows have approximately the same number of dots. In the membership table in Figure 5, all the column have exactly the same numbers of dots and all the rows have exactly the same number of dots. For this reason, we say that

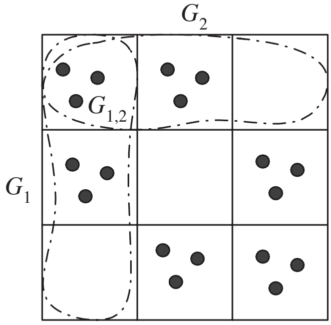

Figure 5: The membership table for a finite group G and subgroups G_1 and G_2.

the table exhibits a *quasi-uniform* structure. In a membership table, each entry can contain a constant number of dots, while in a strong typicality array, each entry can contain only one dot.

Theorem 3. *Let $G_i, i \in [n]$ be subgroups of a group G. Then $\mathbf{h} \in \mathcal{H}_n$ defined by*

$$h_\alpha = \log \frac{|G|}{|G_\alpha|}$$

for all $\alpha \in \overline{\mathsf{N}}$ is entropic, i.e., $\mathbf{h} \in \Gamma_n^*$.

Proof. It suffices to show that there exists a collection of random variables $X_1, X_2, \ldots, X_n$ such that

$$H(X_\alpha) = \log \frac{|G|}{|G_\alpha|} \tag{1}$$

for all $\alpha \in \overline{\mathsf{N}}$. We first introduce a uniform random variable Λ defined on the sample space G with probability mass function

$$\Pr\{\Lambda = a\} = \frac{1}{|G|}$$

for all $a \in G$. For any $i \in [n]$, let random variable X_i be a function of Λ such that $X_i = aG_i$ if $\Lambda = a$.

Consider any $\alpha \in \overline{\mathsf{N}}$. Since $X_i = a_i G_i$ for all $i \in \alpha$ if and only if Λ is equal to some $b \in \cap_{i\in\alpha} a_i G_i$,

$$\begin{aligned}\Pr\{X_i = a_i G_i : i \in \alpha\} &= \frac{|\bigcap_{i\in\alpha} a_i G_i|}{|G|} \\ &= \begin{cases} \frac{|G_\alpha|}{|G|} & \text{if } \bigcap_{i\in\alpha} a_i G_i \neq \emptyset \\ 0 & \text{otherwise} \end{cases}\end{aligned}$$

by Lemma 1. In other words, $(X_i, i \in \alpha)$ is distributed uniformly on its support whose cardinality is $\frac{|G|}{|G_\alpha|}$. Then (1) follows and the theorem is proved. □

This theorem shows that an entropy function for n random variables $X_1, X_2, \ldots, X_n$ can be constructed from any finite group G and subgroups $G_1, G_2, \ldots, G_n$, with

$$H(X_\alpha) = \log \frac{|G|}{|G_\alpha|}, \quad \alpha \in \overline{\mathsf{N}},$$

which depends only on the orders of G and $G_1, G_2, \ldots, G_n$. Now consider the entropy inequality

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

that holds for all random variables X_1 and X_2, in particular for X_1 and X_2 constructed from any finite group G and subgroups G_1 and G_2 by means of Theorem 3. Substituting this entropy function into the inequality, we obtain

$$\log \frac{|G|}{|G_1|} + \log \frac{|G|}{|G_2|} \geq \log \frac{|G|}{|G_1 \cap G_2|}, \tag{2}$$

or

$$|G||G_1 \cap G_2| \geq |G_1||G_2|.$$

This group inequality is well-known in group theory and can be proved by group theoretic means (see for example [64, Sec. 16.4]).

The non-Shannon-type inequality ZY98, expressed in joint entropies, has the form

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

From this, we can obtain the group inequality

$$\left.\begin{array}{r}|G_1 \cap G_3|^3 |G_1 \cap G_4|^3 \\ \cdot\, |G_3 \cap G_4|^3 |G_2 \cap G_3| \\ \cdot\, |G_2 \cap G_4|\end{array}\right\} \le \left\{\begin{array}{l}|G_1||G_1 \cap G_2||G_3|^2 \\ \cdot\, |G_4|^2 |G_1 \cap G_3 \cap G_4|^4 \\ \cdot\, |G_2 \cap G_3 \cap G_4|\end{array}\right. ,$$

which can be called a "non-Shannon-type" group inequality. To our knowledge, there has not been a group theoretic proof of this inequality.

Hence, for any entropy inequality that holds for any n random variables, one can obtain a corresponding inequality that holds for any finite group and any n of its subgroups. It can be shown that for any group inequality of the form (2) that holds for any finite group and any n of its subgroups, the corresponding entropy inequality also holds for any n random variables. This establishes a one-to-one correspondence between entropy and finite groups.

3.3. Probability theory

In probability theory, a central notion is *conditional independence* of random variables. The relation between conditional independence and constraints on the entropy function is the following: For $\alpha, \beta, \gamma \in \mathsf{N}$, X_α and X_β independent conditioning on X_γ if and only if $I(X_\alpha; X_\beta | X_\gamma) = 0$.

We write "$X_\alpha \perp X_\beta | X_\gamma$" for the conditional independency (CI) $I(X_\alpha; X_\beta | X_\gamma) = 0$. Since $I(X_\alpha; X_\beta | X_\gamma) = 0$ is equivalent to

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

"$X_\alpha \perp X_\beta | X_\gamma$" corresponds to the hyperplane

$$\{\mathbf{h} \in \mathcal{H}_n : h_{\alpha\cup\gamma} + h_{\beta\cup\gamma} - h_{\alpha\cup\beta\cup\gamma} - h_\gamma = 0\}.$$

For a CI K, we denote the hyperplane in $\mathcal{H}_n$ corresponding to K by $\mathcal{E}(K)$. For a collection $\Pi = \{K\}$ of CIs, with a slight abuse of notation, let $\mathcal{E}(\Pi) = \cap_{K\in\Pi}\mathcal{E}(K)$. Then a collection of random variables Θ satisfies Π if and only if $\mathbf{h}^\Theta \in \mathcal{E}(\Pi)$. This gives a geometrical interpretation for conditional independence.

The relation between conditional independence and constraints on the entropy function does not stop here. In probability problems, we are often given a set of CI's and we need to determine whether another given CI is logically implied. This problem, called the *implication problem*, is one of the most basic problems in probability theory. As an example, consider random variables X_1, X_2, and X_3 that satisfy "$X_1 \perp X_3 | X_2$" and "$X_1 \perp X_2$". Then we have

$$\begin{aligned}
0 &\leq I(X_1; X_3) \\
&= I(X_1; X_2, X_3) - I(X_1; X_2 | X_3) \\
&= I(X_1; X_2) + I(X_1; X_3 | X_2) - I(X_1; X_2 | X_3) \\
&= 0 + 0 - I(X_1; X_2 | X_3) \\
&= -I(X_1; X_2 | X_3) \\
&\leq 0,
\end{aligned}$$

where we have invoked two basic inequalities. Therefore, $I(X_1; X_3) = 0$, and we have shown that

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

This example shows that certain structure of conditional independence can be implied by constraints on the entropy functions. In fact, the complete structure of conditional independence is implied by constraints on the entropy functions, namely through the characterization of the region Γ_n^*. To explain this, we first need to explain the building blocks of conditional independence for n random variables $X_i, i \in [n]$. It can be shown that every Shannon information measure involving $X_i, i \in [n]$ can be expressed as the sum of Shannon information measures of the following two *elemental forms*:

i) $H(X_i | X_{[n]-\{i\}}), i \in [n]$;

ii) $I(X_i; X_j | X_K)$, where $i \neq j$ and $K \subset [n] - \{i, j\}$.

For example, it can easily be verified that

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

where the right hand side consists of elemental Shannon information measures for $n = 3$. Then the basic inequality $H(X_1, X_2) \geq 0$ can be obtained by summing (is implied by) the corresponding elemental inequalities:

$$\begin{aligned} H(X_1|X_2,X_3) &\geq 0 \\ I(X_1;X_2) &\geq 0 \\ I(X_1;X_3|X_2) &\geq 0 \\ H(X_2|X_1,X_3) &\geq 0 \\ I(X_2;X_3|X_1) &\geq 0. \end{aligned}$$

This is the reason for the name "elemental inequalities", because for a fixed n, the basic inequalities are implied by the subset of elemental inequalities.

For a fixed n, by setting the two forms of elemental Shannon information measures to 0, we obtain the corresponding forms of *elemental conditional independencies*. Note that the first elemental form, namely $H(X_i|X_{[n]-\{i\}})$, can be written as $I(X_i;X_i|X_{[n]-\{i\}})$, and so $H(X_i|X_{[n]-\{i\}}) = 0$ (a functional dependency) is regarded as a special case of conditional independency.

We now explain why it suffices to consider all elemental conditional independencies (ECIs) instead of all conditional independencies (CIs) that involve $X_i, i \in [n]$. As an example, fix $n = 3$ and consider

$$I(X_1,X_2;X_3) = I(X_2;X_3) + I(X_1;X_3|X_2).$$

Since both $I(X_2;X_3)$ and $I(X_1;X_3|X_2)$ are nonnegative, $I(X_1,X_2;X_3)$ vanishes if and only if both $I(X_2;X_3)$ and $I(X_1;X_3|X_2)$ vanish. Therefore, the CI "$(X_1,X_2) \perp X_3$" is equivalent to the ECIs "$X_2 \perp X_3$" and "$X_1 \perp X_3|X_2$". Therefore, ECIs are the building blocks of the structure of conditional independence of random variables.

The compatibility of ECIs has been studied systematically by Matúš and Studený [46] and Matúš [42, 43] (specifically for $n = 4$), in which the p-representability problem was formulated as follows. Let $\mathrm{ECI}(n)$ denote the collection of all ECIs for any collection Θ of n random variables. Let $\{A, A^c\}$ denote a partition of $\mathrm{ECI}(n)$ (either A or A^c may be empty). Then for any $A \subset \mathrm{ECI}(n)$, we ask whether there exists a particular Θ such that Θ satisfies all $K \in A$ but does not satisfy any $K \in A^c$. If so, we say that $\{A, A^c\}$ is p-representable, otherwise we say that $\{A, A^c\}$ is not p-representable.

The problem of characterizing Γ_n^* subsumes the p-representability problem; the latter completely captures the structure of conditional independence of random variables. Specifically, $\{A, A^c\}$ is p-representable if and only if

$$\exists\, \mathbf{h} \in \Gamma_n^* \text{ s.t. } \mathbf{h} \in \mathcal{E}(A) \backslash \mathcal{E}(A^c),$$

or equivalently,

$$\Gamma_n^* \cap \left(\bigcap_{K \in A} \mathcal{E}(K) \right) \Big\backslash \left(\bigcap_{K \in A^c} \mathcal{E}(K) \right) \neq \emptyset.$$

Note that it takes more than a characterization of $\overline{\Gamma}_n^*$ to solve the p-representability problem.

The p-representability problem in turn subsumes the implication problem. Specifically, a collection of CIs $\Pi \subset \mathrm{ECI}(n)$ implies a CI $K \in \mathrm{ECI}(n)$ if and only if

$$\forall\, \mathbf{h} \in \Gamma_n^*, \mathbf{h} \in \mathcal{E}(\Pi) \Rightarrow \mathbf{h} \in \mathcal{E}(K),$$

or equivalently,

$$\Gamma_n^* \cap \mathcal{E}(\Pi) \subset \mathcal{E}(K).$$

The implication problem and hence the p-representability problem are surprising difficult. It was not until the late 1990's that Matúš [43] settled the p-representability problem for $n = 4$ by first establishing a constrained non-Shannon-type inequality which is a variation of ZY97. The general problem is still open. The special case of the problem for any n when all the CIs are *full conditional independencies*[6] (FCIs) was solved by Yeung et al. [67]. In particular, a Markov random field can be specified as a collection of FCIs. The characterization of the structure of full conditional independence is purely graph theoretic and is implied by Shannon-type inequalities.

3.4. 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 respective 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 [35] for a comprehensive treatment of the subject.

Hammer et al. [26] have shown that all linear inequalities that are valid for Kolmogorov complexity are also valid for entropy, and vice versa. For

[6] A CI "$X_\alpha \perp X_\beta | X_\gamma$" is an FCI for a given n if $\{\alpha, \beta, \gamma\}$ is a partition of $[n]$ (α, β, and γ not necessarily nonempty).

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.

3.5. Network coding

For a long time, information transmission in a point-to-point network had been by and large regarded as commodity flow in the network, where routing is the only operation performed by the intermediate nodes. In the 1970's, Celebiler and Stette [8] proposed the use of coding in a satellite communication system for the purpose of improving the downlink capacity when the ground stations are considered in pairs.[7] Instead of broadcasting the data streams of the two ground stations separately, the modulo 2 sum of the two data streams are broadcast. This work, inspired by Shannon's work on the two-way channel [57], first proposed the use of coding at an intermediate node of a network.

In the 1990's, Yeung [62] studied a distributed data storage problem and discovered that unlike point-to-point communication, in network communication, joint coding of independent information sources is sometimes necessary in order to achieve the network capacity.[8] This was indeed the case for the satellite communication system studied in [8] although it was not explicitly discussed therein. Subsequently, Yeung and Zhang [69] considered the more general satellite communication problem in which multiple ground stations multicast different information sources to different sets of

[7] The author would like to thank Prof. Don Towsley for pointing out this reference.

[8] Consider two independent information sources X and Y to be transmitted in a point-to-point communication system. If we compress X and Y jointly, we need to transmit approximately $H(X, Y)$ bits. If we compress X and Y separately, we need to transmit approximately $H(X) + H(Y)$ bits. But since X and Y are independent, we have $H(X, Y) = H(X) + H(Y)$. Roughly speaking, joint coding of independent information sources is not necessary in a point-to-point communication system.

ground stations. In Ahlswede et al. [1], the advantage of network coding over routing was explicitly demonstrated by an example now known as *the butterfly network*,[9] and the term "network coding", which refers to coding at the intermediate nodes of a network, was coined. In this work, they studied *single-source network coding* in which a single information source is multicast from a source node to a set of sink nodes in a general point-to-point network.

It was established in [1] that the network capacity for single-source network coding admits a simple graph theoretic characterization in the form of a max-flow min-cut theorem for information flow that generalizes the corresponding classical theorem for commodity flow [21, 22]. Subsequently, it was proved by Li et al. [36] and then by Koetter and Médard [34] that linear network coding suffices to achieve the network capacity.

However, for the more general problem with multiple information sources, characterization of the network capacity is much more difficult. In [69], inner and outer bounds on the network capacity in terms of the region of entropy functions, i.e., Γ^*, were obtained. This work was further developed into *multi-source network coding* by Song et al. [58], in which the multi-source multicast problem was considered on general acyclic networks. An exact characterization of the capacity for multi-source network coding (for acyclic networks) was finally obtained by Yan et al. [60, 61]. However, this characterization is implicit in the sense that it is not computable, precisely because the determination of Γ^* remains open.

Dougherty et al. [20] discovered the first example of multi-source network coding whose capacity characterization requires the use of ZY98. Chan and Grant [12] obtained a duality between entropy functions and network coding which asserts that for every $\mathbf{h} \in \mathcal{H}_n$, there exists a multi-source network coding problem characterized by $\mathbf{h}$, such that the problem has a network solution if and only if $\mathbf{h} \in \overline{\Gamma}^*_n$.

The insufficiency of specific forms of linear coding for multi-source network coding were demonstrated and discussed by Riis [55], Rasala Lehman and Lehman [54], and Médard et al. [47]. The insufficiency of very general forms of linear coding has been proved by Dougherty et al. [18]. This is also implied by the result of Chan and Grant [12], because compared with the entropy function, the rank function of vector spaces satisfies additional constraints, in particular the Ingleton inequality [33].

The theory of linear network coding has been generalized to *network error correction* by Yeung and Cai [6, 66] and *secure network coding* by Cai

[9] The name was coined by Michelle Effros.

and Yeung [7]. Along a related line, Beimel et al. [3] have applied ZY98 to obtain new performance bounds in secret sharing, which can be regarded as a special case of secure network coding. An interpretation of secret sharing problems in terms of Γ^* and Γ can be found in [44, Section IV].

In quantum information theory, *quantum network coding* has been studied by Hayashi et al. [31].

3.6. Matrix theory

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

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

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

$$h(\mathbf{X}) = -\int f(\mathbf{x}) \log f(\mathbf{x}) d\mathbf{x}. \tag{3}$$

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

A linear differential entropy inequality

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

is said to be balanced if for all $i \in [n]$, we have $\sum_{\alpha \in \overline{\mathsf{N}} : i \in \alpha} c_\alpha = 0$. (The same can be defined for an entropy inequality.) Chan [10] showed 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 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 (3), we obtain

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

where $|\cdot|$ denotes the determinant of a matrix. For $\alpha \in \overline{\mathsf{N}}$, 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 $\mathbf{X}_\alpha$ is also Gaussian, it follows that

$$h(\mathbf{X}_\alpha) = \frac{1}{2} \log \left[(2\pi e)^{|\alpha|} |K_\alpha|\right]. \tag{4}$$

Now consider the independence bound for differential entropy,

$$h(X_1, X_2, \ldots, X_n) \le \sum_i h(X_i),$$

which is tight if and only if $X_i, i \in [n]$ are mutually independent. Substituting (4) into the above, we have

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

or

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

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

$$|K| \le \prod_i K_i,$$

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 $i \ne j$.

For every valid differential entropy inequality, a corresponding inequality involving the principal minors of a positive 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 definite matrix K. Recently, Chan et al. [13] showed that for 3 × 3 positive definite matrices, all inequalities involving the principal minors can be obtained through the Gaussian distribution as explained. In a related work, Hassibi and Shadbakht [30] studied the properties of normalized Gaussian (differential) entropy functions.

3.7. 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

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

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 [2]), is regarded as the quantum analog of the

[10] We refer the reader to the book by Nielsen and Chuang [50] for quantum information theory.

inequality

$$H(X) \leq H(X,Y) \tag{5}$$

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 (5), 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 proved by Lieb and Ruskai [37, 38] 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* [49]:

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

Following the discovery of non-Shannon-type inequalities for the classical entropy, it became natural to ask whether there exist constraints on the

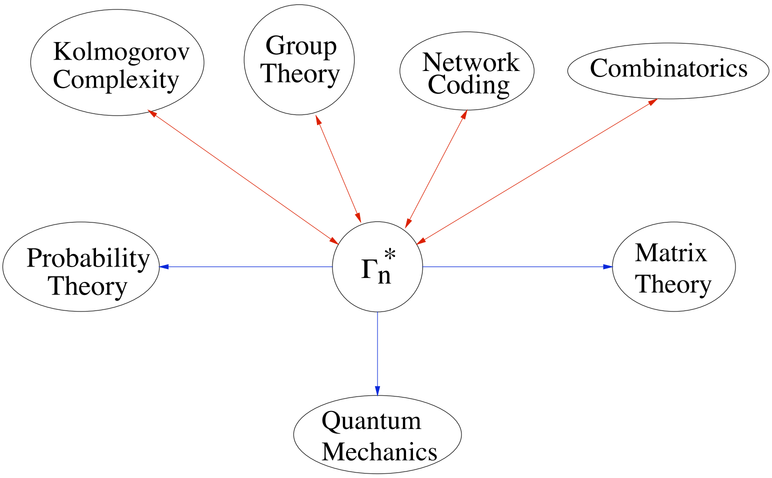

Figure 6: Connections between entropy and various subjects in information sciences, mathematics, and physics.

von Neumann entropy beyond strong subadditivity. Pippenger [52] proved that for a three-party system, there exist no such constraint. Subsequently, Linden and Winter [39] discovered for a four-party system a constrained inequality for the von Neumann entropy which is independent of strong subadditivity. Recently, Cadney et al. [5] proved a family of countably infinitely many constrained inequalities that are independent of each other and strong subadditivity.

4. Concluding remarks

We have presented a comprehensive discussion on the connections between entropy and a number of seemingly unrelated subjects in information sciences, mathematics, and physics. These connections are summarized in the diagram in Figure. 6. In this diagram, Γ_n^*, denoting entropy, is connected by double arrows with combinatorics, group theory, Kolmogorov complexity, and network coding, meaning that there is a one-to-one correspondence for each of these pairs. This suggest the existence of a common underlying structure for all these five subjects. The exact relations among these subjects are still highly confounding, although the quasi-uniform array appears to play a central role in these relations.

With the one-to-one correspondence between entropy and finite groups, we have seen how the rich set of tools in information theory can be employed to obtain results in group theory. The other research direction is less explored but is potentially very fertile. The same can be said for the one-to-one correspondence between entropy and Kolmogorov complexity.

In the same diagram, Γ_n^* is connected by single arrows to probability theory, matrix theory, and quantum mechanics. The studies of entropy have made direct impacts on probability theory and matrix theory. For quantum mechanics, inspirations from classical information theory have borne fruits in quantum information theory.

This expository work does not aim to draw a conclusion on all the findings discussed here. Rather, it serves as a preamble to a series of investigations that will keep researchers from different fields busy for a very long time.

Acknowledgments

The author would like to thank Dr. Siu Wai Ho for contributing Section 3.7 on quantum mechanics, and Dr. Terence Chan, Dr. Fan Cheng, and Mr. Qi Chen for the useful discussions.

References

[1] R. Ahlswede, N. Cai, S.-Y. R. Li and R. W. Yeung, *Network information flow.* IEEE Trans. Info. Theory, **IT-46** (2000), 1204–1216.

[2] H. Araki and E. H. Lieb. *Entropy inequalities.* Comm. Math. Phys., **18** (1970), 160–170.

[3] A. Beimel, N. Livne and C. Padro, *Matroids can be far from ideal secret sharing.* Proc. of the Fifth Theory of Cryptography Conference, New York, NY, Mar. 19–21, 2008.

[4] T. Berger, *Multiterminal source coding.* In: The Information Theory Approach to Communications (G. Longo ed.), CISM Courses and Lectures #229, Springer-Verlag, New York, 1978.

[5] J. Cadney, N. Linden and A. Winter, *Infinitely many constrained inequalities for the von Neumann entropy.* IEEE Trans. Info. Theory, **IT-58** (2012), 3657–3663.

[6] N. Cai and R. W. Yeung, *Network error correction, Part II: Lower bounds.* Comm. Info. and Syst., **6** (2006), 37–54.

[7] N. Cai and R. W. Yeung, *Secure network coding on a wiretap network.* IEEE Trans. Info. Theory, **IT-57** (2011), 424–435.

[8] M. Celebiler and G. Stette, *On increasing the down-link capacity of a regenerative satellite repeater in point-to-point communications.* Proceedings of the IEEE, **66** (1978), no. 1, 98–100.

[9] T. H. Chan, *A combinatorial approach to information inequalities.* Comm. Info. and Syst., **1** (2001), 241–253.

[10] T. H. Chan, *Balanced information inequalities.* IEEE Trans. Info. Theory, **IT-49** (2003), 3261–3267.

[11] T. Chan, *Recent progresses in characterizing information inequalities.* Entropy, **13** (2011), 379–401.

[12] T. Chan and A. Grant, *Dualities between entropy functions and network codes.* IEEE Trans. Info. Theory, **IT-54** (2008), 4470–4487.

[13] T. 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.

[14] T. H. Chan and R. W. Yeung, *On a relation between information inequalities and group theory.* IEEE Trans. Info. Theory, **IT-48** (2002), 1992–1995.

[15] S.-Y. Chung, *Information-Theoretic Theorem Prover.* `http://itl.kaist.ac.kr/ittp.html`, 2009.

[16] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, Wiley, 1991, 2nd ed., Wiley-Interscience, 2006.

[17] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems.* Academic Press, New York, 1981.

[18] R. Dougherty, C. Freiling and K. Zeger, *Insufficiency of linear coding in network information flow.* IEEE Trans. Info. Theory, **IT-51** (2005), 2745–2759.

[19] R. Dougherty, C. Freiling and K. Zeger, *Six new non-Shannon information inequalities.* 2006 IEEE International Symposium on Information Theory, Seattle, WA, Jul. 9–14, 2006.

[20] R. Dougherty, C. Freiling and K. Zeger, *Networks, matroids, and non-Shannon information inequalities.* IEEE Trans. Info. Theory, **IT-53** (2007), 1949–1969.

[21] P. Elias, A. Feinstein and C. E. Shannon, *A note on maximum flow through a network.* IRE Trans. Info. Theory, **IT-2** (1956), 117–119.

[22] L. R. Ford, Jr. and D. R. Fulkerson, *Maximal flow through a network.* Canadian J. of Math., **VIII** (1956), 399–404.

[23] S. Fujishige, *Polymatroidal dependence structure of a set of random variables.* Info. Contr., **39** (1978), 55–72.

[24] R. G. Gallager, *Information Theory and Reliable Communication*, Wiley, New York, 1968.

[25] J. Dj. Golić, *Noiseless coding for multiple channels.* 1994 International Symposium on Information Theory and Its Applications, Sydney, Australia, 1994.

[26] D. Hammer, A. Romashchenko, A. Shen and N. Vereshchagin, *Inequalities for Shannon Entropy and Kolmogorov Complexity.* J. Comp. and Syst. Sci., **60** (2000), 442–464.

[27] T. S. Han, *Linear dependence structure of the entropy space.* Info. Contr., **29** (1975), 337–368.

[28] T. S. Han, *A uniqueness of Shannon's information distance and related non-negativity problems.* J. Comb., Info., and Syst. Sci., **6** (1981), 320–321.

[29] B. Hassibi and S. Shadbakht, *Normalized entropy vectors, network information theory and convex optimization.* 2007 IEEE Information Theory Workshop on Information Theory for Wireless Networks, Bergen, Norway, Jul 1–6, 2007.

[30] B. Hassibi and S. Shadbakht, *The entropy region for three Gaussian random variables.* 2008 IEEE International Symposium on Information Theory, Toronto, Canada, Jul 6–11, 2008.

[31] M. Hayashi, K. Iwama, H. Nishimura, R. Raymond and S. Yamashita, *Quantum network coding.* Lecture Notes in Comp. Sci., **4393** (2007), 610–621.

[32] S. W. Ho and R. W. Yeung, *On information divergence measures and a unified typicality.* IEEE Trans. Info. Theory, **IT-56** (2010), 5893–5905.

[33] A. W. Ingleton, *Representation of matroids.* In: Combinatorial Mathematics and Its Applications (D. J. A. Welsh ed.), 149–167, Academic Press, London, 1971.

[34] R. Koetter and M. Médard, *An algebraic approach to network coding.* IEEE/ACM Trans. Networking, **11** (2003), 782–795.

[35] M. Li and P. Vitányi, *An Introduction to Kolmogorov Complexity and Its Applications.* 3rd ed., Springer, New York, 2008.

[36] S.-Y. R. Li, R. W. Yeung and N. Cai, *Linear network coding.* IEEE Trans. Info. Theory, **IT-49** (2003), 371–381.

[37] E. H. Lieb and M. B. Ruskai, *A fundamental property of quantum-mechanical entropy.* Phys. Rev. Lett., **30** (1973), no. 10, 434–436.

[38] E. H. Lieb and M. B. Ruskai, *Proof of the strong subadditivity of quantum mechanical entropy.* J. Math. Phys., **14** (1973), 1938–1941.

[39] N. Linden and A. Winter, *A new inequality for the von Neumann entropy.* Comm. Math. Phys., **259** (2005), 129–138.

[40] K. Makarychev, Y. Makarychev, A. Romashchenko and N. Vereshchagin, *A new class of non-Shannon-type inequalities for entropies.* Comm. Info. and Syst., **2** (2002), 147–166.

[41] F. Matúš, *Probabilistic conditional independence structures and matroid theory: Background.* Int. J. of General Syst., **22** (1994), 185–196.

[42] F. Matúš, *Conditional independences among four random variables II.* Combinatorics, Probability and Computing, **4** (1995), 407–417.

[43] F. Matúš, *Conditional independences among four random variables III: Final conclusion.* Combinatorics, Probability and Computing, **8** (1999), 269–276.

[44] F. Matúš, *Two constructions on limits of entropy functions.* IEEE Trans. Info. Theory, **IT-53** (2007), 320–330.

[45] F. Matúš, *Infinitely many information inequalities.* 2007 IEEE International Symposium on Information Theory, Nice, France, Jun. 24–29, 2007.

[46] F. Matúš and M. Studený, *Conditional independences among four random variables I.* Combinatorics, Probability and Computing, **4** (1995), 269–278.

[47] M. Médard, M. Effros, T. Ho and D. Karger, *On coding for nonmulticast networks.* 41st Annual Allerton Conference on Communication, Control, and Computing, Monticello, IL, Oct. 2003.

[48] J. von Neumann, *Mathematische Grundlagen der Quantenmechanik.* Springer, Berlin, 1932.

[49] M. A. Nielsen.*Quantum Information Theory.* Ph.D. thesis, University of New Mexico, 1998.

[50] M. A. Nielsen and I. L. Chuang, *Quantum Computation and Quanum Information.* Cambridge University Press, 2000.

[51] N. Pippenger, *What are the laws of information theory?* 1986 Special Problems on Communication and Computation Conference, Palo Alto, CA, Sept. 3–5, 1986.

[52] N. Pippenger, *The inequalities of quantum information theory.* IEEE Trans. Info. Theory, **IT-49** (2003), 773–789.

[53] R. Pulikkoonattu, E. Perron and S. Diggavi, *Xitip.* `http://xitip.epfl.ch`, 2008.

[54] A. Rasala Lehman and E. Lehman, *Complexity classification of network information flow problems.* 41st Annual Allerton Conference on Communication, Control, and Computing, Monticello, IL, Oct. 2003.

[55] S. Riis, *Linear versus nonlinear boolean functions in network flows.* 38th Annual Conference on Information Sciences and Systems (CISS), Princeton, NJ, Mar. 17–19, 2004.

[56] C. E. Shannon, *A Mathematical Theory of Communication.* Bell Sys. Tech. Journal, **27** (1948), 379–423, 623–656.

[57] C. E. Shannon, *Two-way communication channels.* Proc. 4th Berkeley Symp. Math. Statist. and Prob., **1** (1961), 611–644.

[58] L. Song, R. W. Yeung and N. Cai, *A separation theorem for single-source network coding.* IEEE Trans. Info. Theory, **IT-52** (2006), 1861–1871.

[59] J. Wolfowitz, *Coding Theorems of Information Theory.* Springer, Berlin-Heidelberg, 2nd ed., 1964, 3rd ed., 1978.

[60] X. Yan, R. W. Yeung and Z. Zhang, *The capacity region for multi-source multi-sink network coding.* 2007 IEEE International Symposium on Information Theory, Nice, France, Jun. 24–29, 2007.

[61] X. Yan, R. W. Yeung and Z. Zhang, *An Implicit Characterization of the Achievable Rate Region for Acyclic Multi-Source Multi-sink Network Coding.* To appear in IEEE Trans. Info. Theory.

[62] R. W. Yeung, *Multilevel diversity coding with distortion.* IEEE Trans. Info. Theory, **IT-41** (1995), 412–422.

[63] R. W. Yeung, *A framework for linear information inequalities.* IEEE Trans. Info. Theory, **IT-43** (1997), 1924–1934.

[64] R. W. Yeung, *A First Course in Information Theory.* Kluwer Academic/Plenum Publishers, New York, 2002.

[65] R. W. Yeung, *Information Theory and Network Coding.* Springer, 2008.

[66] R. W. Yeung and N. Cai, *Network error correction, Part I: Basic concepts and upper bounds.* Comm. Info. and Syst., **6** (2006), 19–36.

[67] R. W. Yeung, T. T. Lee and Z. Ye, *Information-theoretic characterization of conditional mutual independence and Markov random fields.* IEEE Trans. Info. Theory, **IT-48** (2002), 1996–2011.

[68] R. W. Yeung and Y.-O. Yan, *Information-Theoretic Inequality Prover (ITIP).* `http://user-www.ie.cuhk.edu.hk/~ITIP/`, 1996.

[69] R. W. Yeung and Z. Zhang, *Distributed source coding for satellite communications.* IEEE Trans. Info. Theory, **IT-45** (1999), 1111–1120.

[70] Z. Zhang, *On a new non-Shannon-type information inequality.* Comm. Info. and Syst., **3** (2003), 47–60.

[71] Z. Zhang and R. W. Yeung, *A non-Shannon-type conditional inequality of information quantities.* IEEE Trans. Info. Theory, **IT-43** (1997), 1982–1986.

[72] Z. Zhang and R. W. Yeung, *On characterization of entropy function via information inequalities.* IEEE Trans. Info. Theory, **IT-44** (1998), 1440–1452.

Institute of Network Coding, Department of Information Engineering, and Shenzhen Research Institute
The Chinese University of Hong Kong
N.T., Hong Kong
E-mail address: `whyeung@ie.cuhk.edu.hk`

Received September 24, 2012
Accepted September 24, 2015