On Entropy, Information Inequalities, and Groups

Raymond W. Yeung
The Chinese University of Hong Kong
whyeung@ie.cuhk.edu.hk

Abstract

There has been significant progress in the study of entropy functions and information inequalities in the past 10 years. The set-theoretic structure of Shannon's information measures has been established, and machine-proving of most information inequalities known to date (Shannon-type inequalities) has become possible. Most importantly, the recent discovery of a few so-called non-Shannon-type inequalities reveals the existence of information inequalities which cannot be proved by techniques known during the first 50 years of information theory. In this expository paper, we explain the essence of this fundamental subject, give a number of applications of the results, and discuss their implications in information theory, probability theory, and group theory.

1 Introduction

Let $\mathcal{N} = \{1, \ldots, n\}$ and let $\Omega = \{X_i, i \in \mathcal{N}\}$ be any collection of n discrete random variables. Associated with $\{X_i, i \in \mathcal{N}\}$ are $2^n - 1$ joint entropies. For example, when $n = 3$, the 7 joint entropies are

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

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

Note that all other types of Shannon's information measures, namely mutual information, conditional entropy, and conditional mutual information, are all linear combinations of (unconditional) joint entropies.

For any subset α of $\mathcal{N}$, let $X_\alpha = (X_i, i \in \alpha)$ and $H_\Omega(\alpha) = H(X_\alpha)$. For fixed Ω, one can then view H_Ω as a set function from $2^\mathcal{N}$ to $\mathbb{R}$ with $H_\Omega(\phi) = 0$, i.e., we adopt the convention that the entropy of an empty set

of random variables is equal to zero. For this reason, we call H_Ω the entropy function of Ω.

It is well-known that for any Ω, H_Ω satisfies the following properties for all $\alpha, \beta \subset \mathcal{N}$:

(P1) $H_\Omega(\phi) = 0$;

(P2) $H_\Omega(\alpha) \leq H_\Omega(\beta)$ if $\alpha \subset \beta$;

(P3) $H_\Omega(\alpha) + H_\Omega(\beta) \geq H_\Omega(\alpha \cap \beta) + H_\Omega(\alpha \cup \beta)$.

These are called the *polymatroidal axioms*. These axioms are equivalent to the nonnegativity of all Shannon's information measures involving random variables in Ω, which are called the *basic inequalities* [20] (see Appendix A for a proof). In the rest of the paper, we refer to inequalities/identities/expressions involving only Shannon's information measures as information inequalities/identities/expressions.

In the 1986 SPOC Conference, Pippenger gave a talk in which he referred to constraints on entropies as the "laws of information theory" [10]. He asked whether there is any constraint on entropy functions in addition to the basic inequalities.

During the past ten years, there has been significant progress in understanding the properties of entropy functions [12]-[23]. The results obtained not only reveal the set-theoretic structure of Shannon's information measures, but also make machine-proving of information inequalities possible. Moreover, owing to the recent discovery of a new inequality [22], Pippenger's open problem is finally settled. In other words, the polymatroidal axioms actually form an incomplete set of constraints on entropy functions. It is now believed that there are many information inequalities which are still unknown. Further investigation along this line may eventually lead us to new territories in information theory and probability theory. In particular, some yet undiscovered inequalities may make the solution of certain open problems in multiusers information theory possible. For a historical note on the subject, we refer the readers to Section 1 in [22].

The purpose of the current paper is to describe the essence of this subject without giving most of the technical details. For a comprehensive treatment of the subject, we refer the reader to [27]. The rest of the paper consists of six sections. In Section 2, the fundamental relation between entropy functions and information inequalities is explained. In Section 3, the use of the software ITIP for proving information inequalities is discussed. Section 4 is an introduction of the theory of I-Measure without measure-theoretic

notations. This theory reveals the set-theoretic structure of Shannon's information measures and renders a graphical representation of information measures called *information diagram*. It is also a basic tool for understanding the properties of entropy functions. Section 5 consists of a number of examples which show how the tools discussed in this paper provide a unified approach to understand a class of information theory problems, and how they trivialize the proofs of some basic results in information theory. Section 6 is a discussion of an intriguing relation between information theory and group theory. Concluding remarks are in Section 7.

2 Entropy Functions and Information Inequalities

Recall that H_Ω is a function from $2^\mathcal{N}$ to $\mathbb{R}$ with $H_\Omega(\phi) = 0$. Let $k = 2^n - 1$. Labeling the coordinates of $\mathbb{R}^k$ by $h_\alpha, \alpha \subset 2^\mathcal{N} \backslash \{\phi\}$, where h_α corresponds to the value of $H_\Omega(\alpha)$, an entropy function H_Ω can be represented by a vector in $\mathbb{R}^k$. On the other hand, a vector $h \in \mathbb{R}^k$ is called *entropic* if h represents the entropy function of some collection of n random variables. Define the following region in $\mathbb{R}^k$:

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

For example, when $n = 3$, the coordinates of $\mathbb{R}^7$ are labeled by $h_1, h_2, h_3, h_{12}, h_{13}, h_{23}, h_{123}$, and Γ_3^* is the region in $\mathbb{R}^7$ of all entropy functions of 3 random variables.

An information inequality (linear or nonlinear) has the form $f(h) \geq 0$, where $f : \mathbb{R}^k \to \mathbb{R}$. We consider non-strict inequalities only because these are usually the inequalities of concern in information theory. For example, the inequality $I(X_1; X_2) \geq 0$ is written as $h_1 + h_2 - h_{12} \geq 0$. Since an information inequality involving n random variables *always holds* if and only if it is satisfied by the entropy function of any collection of n random variables, we have the following geometric interpretation of an information inequality:

$$f(h) \geq 0 \text{ always holds if and only if } \Gamma_n^* \subset \{h \in \mathbb{R}^k : f(h) \geq 0\} \ .$$

The two possible cases for $f(h) \geq 0$ are illustrated in Fig. 1 and Fig. 2. Note that Γ_n^* obviously contains the origin, the entropy function of the collection of n degenerate random variables. In Fig. 1, Γ_n^* is completely included by the region $\{h \in \mathbb{R}^k : f(h) \geq 0\}$, so $f(h) \geq 0$ always holds. In Fig. 2, there exists a vector h_0 which corresponds to some entropy function H_Ω such that $f(h_0) < 0$. Thus the inequality $f(h) \geq 0$ does not always hold. If Γ_n^* is

known, we in principle can determine whether any information inequality always holds.

In information theory, we very often deal with information inequalities with certain constraints on the random variables involved. These are called constrained information inequalities. Such constraints on the random variables can usually be expressed as linear constraints on the entropies. The following are such examples:

1. X_1, X_2 and X_3 are mutually independent if and only if $H(X_1, X_2, X_3) = H(X_1) + H(X_2) + H(X_3)$.
2. X_1 is a function of X_2 if and only if $H(X_1|X_2) = 0$.
3. The Markov chain $X_1 \to X_2 \to X_3 \to X_4$ is equivalent to $I(X_1; X_3|X_2) = 0$ and $I(X_1, X_2; X_4|X_3) = 0$.

It turns out that Γ_n^* not only characterizes all unconstrained information inequalities, but also all constrained information inequalities. This is seen by observing that each linear constraint on the entropies is a hyperplane in $\mathbb{R}^k$. In general, linear constraints on the entropies can be expressed as a set of homogeneous linear equations $Qh = 0$. Defining the linear subspace

$$\Phi = \{h \in \mathbb{R}^k : Qh = 0\}$$

and generalizing our interpretation of unconstrained information inequalities, we have

> Under the constraint $Qh = 0$, $f(h) \geq 0$ always holds if and only if $(\Gamma_n^* \cap \Phi) \subset \{h \in \mathbb{R}^k : f(h) \geq 0\}$.

An information identity $f(h) = 0$ always holds if and only if both $f(h) \geq 0$ and $f(h) \leq 0$ always hold. Then we have the following interpretation of a constrained information identity:

> Under the constraint $Qh = 0$, $f(h) = 0$ always holds if and only if $(\Gamma_n^* \cap \Phi) \subset \{h \in \mathbb{R}^k : f(h) = 0\}$.

Unfortunately, Γ_n^* is extremely difficult to characterize, and only partial characterizations of the region have been possible. Let us now define Γ_n as the set of all $h \in \mathbb{R}^k$ which satisfy the following properties for all $\alpha, \beta \subset \mathcal{N}$:

1. $h_\alpha \leq h_\beta$ if $\phi \neq \alpha \subset \beta$;
2. $h_\alpha + h_\beta \geq h_{\alpha\cap\beta} + h_{\alpha\cup\beta}$.

These are precisely the polymatroidal axioms except that the coordinate h_ϕ is degenerated since $H_\Omega(\phi)$ is taken to be 0. Note that Γ_n is also the set of all vectors in $\mathbb{R}^k$ which satisfy the basic inequalities. Since the basic inequalities are observed by all entropy functions, we immediately see that Γ_n is an outer bound on Γ_n^*. The question is whether this outer bound is tight. It turns out that $\Gamma_2^* = \Gamma_2$, but for $n \geq 3$, $\Gamma_n^* \neq \Gamma_n$. In fact, it has been found that Γ_3^* is not even closed [21]!

As Γ_n^* cannot be fully characterized, a more manageable task is to characterize $\overline{\Gamma}_n^*$, the closure of Γ_n^*. If one is interested in unconstrained linear inequalities, then it suffices to consider $\overline{\Gamma}_n^*$ because $\Gamma_n^* \subset \{h \in \mathbb{R}^k : f(h) \geq 0\}$ if and only if $\overline{\Gamma}_n^* \subset \{h \in \mathbb{R}^k : f(h) \geq 0\}$. This follows from the fact that the region $\{h \in \mathbb{R}^k : f(h) \geq 0\}$ is closed. However, if one is interested in constrained inequalities, a more detailed characterization of Γ_n^* is necessary.

It was proved in [21] that $\overline{\Gamma}_n^*$ is in general a convex cone. In the same paper, it was further proved that $\overline{\Gamma}_3^* = \Gamma_3$ (also see [16]). In other words, every unconstrained inequality involving 3 random variables can be proved by invoking the basic inequalities.

In 1998, a new constraint on all entropy functions involving 4 random variables was discovered [22]:

$$I(Z;U) - I(Z;U|X) - I(Z;U|Y)$$

$$\leq \frac{1}{2}I(X;Y) + \frac{1}{4}[I(X;Z,U) + I(Y;Z,U)]. \tag{1}$$

Write the above inequality as $g(h) \geq 0$. Such an inequality is called a *non-Shannon-type inequality* because it is not a consequence of the basic inequalities, i.e.,

$$\Gamma_4 \not\subset \{h \in \mathbb{R}^{15} : g(h) \geq 0\}.$$

On the other hand, since $g(h) \geq 0$ always holds,

$$\Gamma_4^* \subset \{h \in \mathbb{R}^{15} : g(h) \geq 0\}.$$

Taking closure on both sides, we have

$$\overline{\Gamma}_4^* \subset \{h \in \mathbb{R}^{15} : g(h) \geq 0\}.$$

Thus we conclude that $\overline{\Gamma}_4^*$ is a proper subset of Γ_4. This is illustrated in Fig. 3. With the discovery of this inequality, Pippenger's open problem is finally settled. It was further shown in [25] that the inequality (1) implies a class of $2^{14} - 1$ non-Shannon-type constrained inequalities, and physical

interpretations of some of these inequalities were given. Most recently, another class of non-Shannon-type constrained inequalities implied by (1) has been identified [30], and a generalization of (1) has been obtained [31].

We now summarize what we know about Γ_n^*. For $n = 2$,

$$\Gamma_2^* = \overline{\Gamma}_2^* = \Gamma_2.$$

For $n = 3$,

$$\Gamma_3^* \subsetneq \overline{\Gamma}_3^* = \Gamma_3.$$

For $n \geq 4$,

$$\Gamma_n^* \subsetneq \overline{\Gamma}_n^* \subsetneq \Gamma_n.$$

In fact, the first non-Shannon-type inequality, which is a constrained inequality, was discovered earlier in [21]. It was shown that if $I(X;Y) = 0$ and $I(X;Y|Z) = 0$[1], then

$$I(X;Y|Z,U) \leq I(X;Y|U) + I(Z;U|X,Y). \tag{2}$$

Since the constraints on the above inequality are obtained by setting two basic inequalities to equality, this inequality means that there is a certain region on the boundary of Γ_4 which is not entropic. Although this inequality is not strong enough to conclude that $\overline{\Gamma}_4^* \neq \Gamma_4$, it did lead to the conjecture that this is actually the case and eventually to the discovery of (1). Subsequent to [21], a variation of (2) was proved in [23], which was instrumental in settling a longstanding open problem on the conditional independence structure for 4 random variables!

For the inequality in (2), if we further impose the constraint $I(X;Y|U) = I(Z;U|X,Y) = 0$, then we immediately have $I(X;Y|Z,U) = 0$. That is, for 4 random variables X, Y, Z, and U, if 1) X and Y are independent, 2) X and Y are independent given Z, 3) X and Y are independent given U, and 4) Z and U are independent given X and Y, then X and Y given Z and U are independent. This is a constraint on conditional independence relations for 4 random variables which cannot be deduced from the basic inequalities.

3 ITIP — Machine-Proving of Information Inequalities

In the past, information theorists have to prove information inequalities by hand. This is done by successive invocations of the basic inequalities. Such

[1] $I(X;Y) = 0$ and $I(X;Y|Z) = 0$ do not imply each other.

a process can be frustrating when a certain inequality cannot be proved, because we simply do not know whether the inequality is incorrect or we just did not invoke the right basic inequality at the right step. (Of course, we now know that there exist non-Shannon-type inequalities which cannot be proved by this method.)

Now we can prove all Shannon-type information inequalities by a software package called ITIP[2] that runs on MATLAB™. Using ITIP is very simple and intuitive. The following examples illustrate how ITIP is used:

1. `>> ITIP('H(X,Y,Z) >= H(X)+H(Y)+H(Z)')`
   `True`

2. `>> ITIP('I(Y;Z) >= I(X;T)', 'I(X;Z|Y) = 0',`
   `I(XY;T|Z) = 0)`
   `True`

3. `>> ITIP('I(Z;U)-I(Z;U|X)-I(Z;U|Y) <=`
   `0.5I(X;Y)+0.25I(X;ZU)+0.25I(Y;ZU)')`
   `Not provable by ITIP`

In the first example, we prove an unconstrained inequality. In the second example, we prove the data processing theorem. The first inequality is what we want to prove, while the second and the third equalities are the constraints that specify the Markov chain $X \to Y \to Z \to T$. In the third example, we attempt to prove the non-Shannon-type inequality in (1). When ITIP returns the clause "Not provable by ITIP," it means that the inequality may be true but it cannot be deduced from the basic inequalities.

We refer the readers to [20] if they are interested in how ITIP works. Basically, the geometrical interpretation of information inequalities presented in the last section allows one to formulate the problem of proving these inequalities as a linear programming problem.

4 I-Measure and Information Diagrams

In information theory, besides entropies, we also deal with other types of Shannon's information measures. In the last two sections, we simply regard the latter as linear combinations of entropies. By doing so, however, we have totally ignored the underlying set-theoretic structure of Shannon's information measures which gives tremendous insight into information theory problems.

[2]Free download at http://user-www.ie.cuhk.edu.hk/~ITIP/.

The set-theoretic structure of Shannon's information measures has long been observed. The seminal work of Hu [3] in 1962 established that for every information-theoretic identity, a corresponding set-theoretic identity can be obtained via the following substitution of symbols:

$$H/I \to \mu \qquad ;\to \cap \qquad ,\to \cup \qquad |\to - \tag{3}$$

where μ denotes a set-additive function which can take negative values (called a signed measure) and '$-$' is defined by $A - B = A \cap B^c$. For example, from

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

we can obtain

$$\mu(X \cup Y - Z) = \mu(X - Z) + \mu(Y - Z) - \mu(X \cap Y - Z).$$

Hu's work was originally published in Russian and it was not widely known in the West until it was referenced in the book by Csiszár and Körner [8]. With hindsight, a probable reason why there was virtually no further development of Hu's work for three decades is that the set-additive function μ was not explicitly specified.

The set-theoretic nature of Shannon's information measures has also been discussed by various authors in the West. The information theory textbooks by Reza [2] and Abramson [4] both include discussions on the use of Venn diagrams to represent the relation among various information measures for 2 and 3 random variables, but no formal justification was given. Since the late 1960's, such diagrams are not found in most textbooks[3].

In this section, we will describe the theory of I-Measure [12] which was developed on Hu's work and exploits the inclusive-exclusive nature of Shannon's measures. To avoid heavy notation, we will discuss the theory for $n = 3$ only. The idea can easily be extended to the general case. In Fig. 4, we show an *information diagram* for random variables X, Y, and Z. In this diagram, $\tilde{X}$ denotes the set variable corresponding to X, etc. The universal set is defined to be $\tilde{X} \cup \tilde{Y} \cup \tilde{Z}$. For this reason, we call such a diagram an information diagram, a special case of a Venn diagram. In Fig. 4, the region marked with an '*' is identified as $I(X;Y|Z)$, and the two regions marked with a '+' together are identified as $H(Z|X)$. All other Shannon's information measures can be identified likewise according to the substitution of symbols in (3).

[3] One exception is Papoulis's textbook on probability [9]. Cover and Thomas [11] also used a Venn diagram representation for 2 random variables.

Now all Shannon's measures involving these random variables are completely specified by the values of the following 7 joint entropies:

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

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

Consider any signed measure μ^* for the information diagram in Fig. 4 such that

$$\mu^*(\tilde{X}) = H(X),\ \mu^*(\tilde{Y}) = H(Y),\ \mu^*(\tilde{Z}) = H(Z),$$

$$\mu^*(\tilde{X} \cup \tilde{Y}) = H(X,Y), \mu^*(\tilde{Y} \cup \tilde{Z}) = H(Y,Z),$$

$$\mu^*(\tilde{X} \cup \tilde{Z}) = H(X,Z),\ \mu^*(\tilde{X} \cup \tilde{Y} \cup \tilde{Z}) = H(X,Y,Z),$$

i.e., we fix the value of μ^* on a union to be the joint entropy of the corresponding set of random variables. In other words, μ^* is consistent with all entropies via the substitution of symbols in (3). The main observation here is that the values of μ^* on the 7 atoms in the information diagram are uniquely determined by the values of μ^* on the 7 unions. In fact, these two sets of values are invertible linear transformation of each other. Thus the signed measure μ^* is unique. Further, it can be shown that this signed measure is consistent with all other types of Shannon's measures via the substitution of symbols in (3).

With this formulation, under the substitution of symbols in (3), μ is no longer unspecified but is equal to μ^*, which is called the I-Measure for the random variables X, Y, and Z. Now we can formally think of Shannon's measures collectively as a measure μ^*. By doing so, the inclusive-exclusive structure of Shannon's measures can be exploited. For example, $I(X;Y|Z)$, which is equal to $\mu^*(\tilde{X} \cap \tilde{Y} - \tilde{Z})$, is a quantity which belongs to both random variables X and Y but not Z. Also, $I(X;Y) - I(X;Y|Z)$, which is not symbolically symmetrical in X, Y, and Z, is in fact so because it is equal to $\mu^*(\tilde{X} \cap \tilde{Y} \cap \tilde{Z})$. Following the substitution of symbol in (3), we denote this quantity by $I(X;Y;Z)$.

Let us now explain why μ^* in general is not nonnegative. This can be seen by considering X and Y being independent and uniform on $\{0,1\}$, and $Z = X + Y \bmod 2$. It is easy to check that $H(X|Y,Z) = 0$, $I(X;Y|Z) = 1$, and $I(X;Y) = 0$. The same are true for all other permutations of X, Y, and Z. The I-Measure for X, Y, Z is shown in Fig. 5. As can be seen, $I(X;Y;Z)$ is negative. In fact, for any three random variables X, Y, and Z, $\tilde{X} \cap \tilde{Y} \cap \tilde{Z}$ is the only atom on which μ^* can take a negative value, because the value of μ^* on any other atom is equal to a Shannon's measure which is always nonnegative.

Many researchers are not comfortable with the fact that $I(X;Y;Z)$ can be negative because they try to interpret this quantity as a measure of information content (which we do not necessarily have to). Unfortunately, this very often leads to the conclusion that quantities like $I(X;Y;Z)$ are not useful at all. Such a conclusion, however, has been proven wrong. In the next section, we will give several examples of applications of information diagrams in which the quantity $I(X;Y;Z)$ plays a crucial role, and the use of such diagrams trivializes the proofs of many basic results in information theory. Also, the I-Measure was an essential tool in the discovery of the non-Shannon-type inequalities reported in [21, 22].

We end this section by applying the theory of I-Measure to prove a very basic result in information theory which is very often overlooked. As we have mentioned, any linear information expression can be expressed as a linear combination of entropies. We call this the *canonical form* of the information expression. A fundamental question is whether the canonical form is unique. To further elaborate the point, we consider expanding $I(X;Y|Z)$ in the following two ways:

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

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

In the above, we have expanded $I(X;Y|Z)$ in two different ways, each via a series of information identities, into two linear combinations of entropies which happen to be identical. In general, there are many different ways to expand a linear information expression into a linear combination of entropies. The question is whether we would end up with the same linear combination of entropies no matter how the expansion is done. If this is false, then there exists a certain linear combination of entropies whose coefficients are not all equal to zero but the expression is equal to zero for all distributions for the sets of random variables involved.

We now show that this is impossible. For simplicity, we confine our discussion to the case of $n = 2$. (For this case, an entropy function is a vector in $\mathbb{R}^3$.) The argument can readily be extended to any n.

We first show that for any nonnegative measure μ, it is possible to construct random variables X_1 and X_2 such that the I-Measure μ^* is equal to μ. Let

$$\mu(\tilde{X}_1 - \tilde{X}_2) = e_1,\ \mu(\tilde{X}_2 - \tilde{X}_1) = e_2,\ \mu(\tilde{X}_1 \cap \tilde{X}_2) = e_3$$

where $e_1, e_2, e_3 \geq 0$. Let U_1, U_2, and U_3 be mutually independent random variables such that

$$H(U_1) = e_1,\ H(U_2) = e_2,\ H(U_3) = e_3.$$

Consider $X_1 = (U_1, U_3)$ and $X_2 = (U_2, U_3)$, as illustrated in Fig. 6. Then it is easy to see that $\mu^* = \mu$.

The above shows that $h \in \mathbb{R}^3$ is entropic whenever

$$h_{1|2} \stackrel{def}{=} h_{12} - h_2 \geq 0, \qquad h_{2|1} \stackrel{def}{=} h_{12} - h_1 \geq 0,$$

$$i_{1;2} \stackrel{def}{=} h_1 + h_2 - h_{12} \geq 0.$$

Note that $h_{1|2}$, $h_{2|1}$, and $i_{1;2}$ correspond to $H(X_1|X_2)$, $H(X_2|X_1)$, and $I(X_1;X_2)$, respectively, which are the values of μ^* on the atoms. With respect to the coordinates $h_{1|2}, h_{2|1}$, and $i_{1;2}$, the above inequalities defines the positive quadrant of $\mathbb{R}^3$, which has a non-zero Lebesgue measure (i.e., a "volume"). Since $h_{1|2}, h_{2|1}$, and $i_{1;2}$ are invertible linear transformation of h_1, h_2, and h_{12}, with respect to the coordinates h_1, h_2, and h_{12}, the above inequalities again defines a region in $\mathbb{R}^3$ with a non-zero measure. Since this region is in Γ_2^*, we conclude that Γ_2^* has a non-zero measure.

It remains to show that with c_1, c_2, and c_{12} not all equal to zero, $c_1h_1 + c_2h_2 + c_{12}h_{12} = 0$ cannot always hold. Using our interpretation of an information identity, the contrary is true if and only if

$$\Gamma_2^* \subset \{h \in \mathbb{R}^3 : c_1h_1 + c_2h_2 + c_{12}h_{12} = 0\}$$

which is impossible because the set $\{h \in \mathbb{R}^3 : c_1h_1 + c_2h_2 + c_{12}h_{12} = 0\}$ is measure zero while Γ_2^* is not. This proves the uniqueness of the canonical form for linear information expressions. Alternative proofs of this result can be found in [5] and [8] (p. 51, Theorem 3.6).

The uniqueness of the canonical form of a linear information expression has the following application. If we want to see whether two linear information expressions are identical, we only need to express them in canonical form. Then these two expressions are identical if and only if they have the same canonical form.

5 Examples of Application

In this section, we will show by examples how the tools discussed in this paper offer a new approach to understand and analyze many basic problems in information theory. In particular, we will see in a few examples that the quantity $I(X;Y;Z)$ in fact plays a crucial role in the arguments. Therefore, we should exploit instead of avoid this quantity.

Example 1 (Shannon's Perfect Secrecy Theorem [1])

We give a new proof for Shannon's Perfect Secrecy Theorem. Let X be the plain text, Y be the cipher text, and Z be the key in a secret key cryptosystem. In order to achieve perfect secrecy, X and Y must be independent, or $I(X;Y) = 0$. Since X can be recovered from Y and Z, $H(X|Y,Z) = 0$. These are the constraints on X, Y, and Z, and we will show that they imply $H(Z) \geq H(X)$. To this end, let $I(X;Y|Z) = a \geq 0$ (see Fig. 7). Since $I(X;Y) = 0$, we see from Fig. 7 that $I(X;Y;Z) = -a$. In comparing $H(X)$ with $H(Z)$, we do not have to consider $I(X;Z|Y)$ and $I(X;Y;Z)$ since they both belong to $H(X)$ and $H(Z)$. Instead, we only have to compare $H(X|Z)$ and $H(Z|X)$. Now since $I(Y;Z) \geq 0$, we see from the diagram that $I(Y;Z|X) \geq a$. Also, $H(Z|X,Y) \geq 0$. Thus we see that $H(Z) \geq H(X)$. Note that the assumptions that $H(Y|X,Z) = 0$, i.e., the cipher text is a function of the plain text and the key, and $I(X;Z) = 0$, i.e., the plain text and the key are independent, are not necessary in the argument. This result has been generalized to the *imperfect secrecy theorem* in [27] (p. 116).

Example 2 (Convexity of Mutual Information)

Let $(X,Y) \sim p(x,y) = p(x)p(y|x)$. We now show that for fixed $p(x)$, $I(X;Y)$ is a convex function of $p(y|x)$.

Let $p_1(y|x)$ and $p_2(y|x)$ be two conditional distributions. Now consider the system in Fig. 8 in which the position of the switch is determined by a random variable Z with $Pr(Z = 1) = \lambda$ and $Pr(Z = 2) = 1 - \lambda$, and Z is independent of X. The switch takes position i if $Z = i$, $i = 1, 2$. In Fig. 9, let $I(X;Z|Y) = a \geq 0$. Since $I(X;Z) = 0$, we see that $I(X;Y;Z) = -a$. Then

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

$$= \lambda I(p(x), p_1(y|x)) + (1-\lambda) I(p(x), p_2(y|x))$$

where $I(p(x), p_i(y|x))$ denotes the mutual information between the input and output of a channel with input distribution $p(x)$ and transition probability $p_i(y|x)$.

Example 3 (Concavity of Mutual Information)

Let $(X, Y) \sim p(x, y) = p(x)p(y|x)$. We now show that for fixed $p(y|x)$, $p(x)$ is a concave function of $p(x)$.

Consider the system in Fig. 10, where the position of the switch is determined by a random variable Z with the same marginal distribution as in the last example. This time, when X is given, Z is independent of Y, or $I(Y; Z|X) = 0$. Since the atom $\tilde{Y} \cap \tilde{Z} - \tilde{X}$ has zero measure, we degenerate it into a point and obtain the information diagram in Fig. 11. As we have mentioned, for 3 random variables X, Y, and Z, the only atom in the information diagram which can have a negative measure is $\tilde{X} \cap \tilde{Y} \cap \tilde{Z}$. From Fig. 11, we see that under the constraint $I(Y; Z|X) = 0$, $I(X; Y; Z) = I(Y; Z) \geq 0$. Thus μ^* is nonnegative. Again from Fig. 11, since $(\tilde{X} \cap \tilde{Y} - \tilde{Z}) \subset (\tilde{X} \cap \tilde{Y})$, by μ^* being nonnegative, we immediately see that

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

$$= \lambda I(p_1(x), p(y|x)) + (1 - \lambda) I(p_2(x), p(y|x)).$$

Example 4 (Data Processing Theorem)

In the last example, $Z \to X \to Y$ form a Markov chain, and μ^* vanishes on the atom $\tilde{Z} \cap \tilde{X}^c \cap \tilde{Y}$ $(= \tilde{Y} \cap \tilde{Z} - \tilde{X})$. One way to "explain" this phenomenon is that when the vertex X is removed in the graph Z—X—Y, the resulting graph is disconnected. This theme has been generalized to Markov chains with a finite length [13] and to finite Markov random fields satisfying the global Markov property [29]. For Markov chains with a finite length, it was further shown in [13] that μ^* is always nonnegative. By suppressing those atoms on which μ^* vanishes, the information diagram for a Markov chain can be represented in two dimensions in the form of a "fan" (i.e., all the circles intersect at one point)[4]. Fig. 12 shows the information diagram for the Markov chain $X \to Y \to Z \to T$. Since $\tilde{X} \cap \tilde{T}$

[4]The term "fan" was coined by Sergio Verdú. The information diagram for a Markov chain can be drawn alternatively as in [27], which can regarded as the upper part of the fan.

is a subset of $\tilde{Y} \cap \tilde{Z}$ and μ^* is nonnegative, we have $\mu^*(\tilde{X} \cap \tilde{T}) \leq \mu^*(\tilde{Y} \cap \tilde{Z})$, or $I(X;T) \leq I(Y;Z)$, which is the data processing theorem.

Example 5 (Feedback Capacity)

Shannon showed that feedback does not increase the capacity of a discrete memoryless channel. Let us now pretend that Shannon's result is yet unproven, and we are going to explore the problem using a systematic approach. Let W be the message to be sent, and X_i and Y_i be the input and the output of the discrete memoryless channel at time i, $i = 1, 2, \ldots, n$. Fig. 13 specifies the dependency structure of these random variables as an acylic directed graph as we now explain. First, we list these random variables as $WX_1Y_1X_2Y_2 \cdots X_nY_n$. This represents the chronological order in which the random variables are generated. When we go down the list and consider each particular random variable, we find that the predecessors of that random variable in Fig. 13 all have appeared earlier on the list. The interpretation of Fig. 13 is that each random variable depends on all the previous random variables on the list only through its predecessors in the figure. In other words,

$$\begin{aligned} & p(wx_1y_1x_2y_2 \cdots x_ny_n) \\ = \; & p(w)p(x_1|w)p(y_1|x_1)p(x_2|wy_1)p(y_2|x_2) \\ & \cdots p(x_n|wy_1 \cdots y_n)p(y_n|x_n). \end{aligned}$$

This can be translated into the following information-theoretic constraints:

$$\begin{aligned} I(W;Y_1|X_1) &= 0 \\ I(X_1;X_2|W,Y_1) &= 0 \\ I(W,X_1,Y_1;Y_2|X_2) &= 0 \\ &\vdots \\ I(X_1,\cdots,X_{n-1};X_n|W,Y_1,\cdots,Y_{n-1}) &= 0 \\ I(W,X_1,\cdots,X_{n-1},Y_1,\cdots,Y_{n-1};Y_n|X_n) &= 0. \end{aligned}$$

We claim that

$$I(W;Y_1,\cdots,Y_n) \leq \sum_{i=1}^{n} I(X_i;Y_i)$$

for all $n \geq 1$, which is the key to proving the desired result.

ITIP cannot handle problems with an indefinite number of random variables. With a fast workstation, ITIP can handle up to 8 or 9 random

variables. So, we are going to explore the problem for the case of $n = 2$. By renaming X_1 as A, Y_1 as B, X_2 as C, and Y_2 as D (since the current version of ITIP only takes capital letters as random variable names), we run ITIP and prove the claim for $n = 2$ as follows:

```
>> ITIP('I(W;B,D) <= I(A;B)+I(C;D)',
'I(W;B|A) = 0', 'I(A;C|B,W) = 0',
'I(D;W,A,B|C) = 0')
True
```

This lends much evidence to the claim for a general n. Upon finding a way to prove the above inequality by hand (which is not too difficult once it is confirmed to be true), we will have obtained enough insight into the structure of the problem which enables us to prove the general case easily.

We can obtain further insight into the problem by trying to prove the following claims by ITIP (under the same set of constraints):

1. $I(W; B, D|A, C) = 0$
2. $I(W; A, C) \geq I(W; B, D)$
3. $I(A, C; B, D) \leq I(A; B) + I(C; D)$.

For each of these claims, ITIP returns the clause "Not provable by ITIP." Here it does not mean that ITIP has disproved these claims. Rather, it means that these claims cannot be proved by invoking the basic inequalities. Nevertheless, so far there are only a few known problems whose solutions rely on non-Shannon-type inequalities, so we may believe "with high confidence" that these claims are incorrect. In fact, it is not difficult to find counterexamples for these claims, so all of them are indeed false.

Example 6 (Secret Sharing)

Recently there has been much interest in deriving information-theoretic bounds for secret sharing problems. The formulation of a secret sharing problem is as follows. Let $\mathcal{P}$ be a finite set whose elements are called the participants of the secret sharing scheme, let S be a secret random variable shared by all participants $p \in P$, and let $\mathcal{A}$ be a subset of $2^{\mathcal{P}}$ such that for all $A, B \in \mathcal{A}$, $A \not\subset B$ and $B \not\subset A$. $\mathcal{A}$ is called the access structure of the scheme, and if $A \in \mathcal{A}$, the subset of participants A is said to be qualified to know the secret S. A secret sharing scheme is a random vector $(S, X_p : p \in P)$ such that for all $A \subset P$, $H(S|X_p, p \in A)$ equals 0 if there exists $B \in \mathcal{A}$

with $B \subset A$, and equals $H(S)$ otherwise. In other words, if $A \in \mathcal{A}$, then the subset of participants A can recover the secret S perfectly, otherwise they know nothing about the secret. The information rate of the scheme is defined by

$$\frac{H(S)}{\max_p H(X_p)}$$

and the average information rate is defined by

$$\frac{|P|H(S)}{\sum_p H(X_p)}.$$

Capocelli *et al* [14] proved the following result which is useful for bounding the information rates of a secret sharing scheme:

> If a, b, c, d are four participants in $\mathcal{P}$ such that $\{a, b\}, \{b, c\}, \{c, d\} \in \mathcal{A}$ and $\{a, c\}, \{a, d\}, \{b\} \notin \mathcal{A}$, then $H(X_b, X_c) \geq 3H(S)$.

We now see how this result can readily be obtained with the help of ITIP. Here the problem is to obtain the maximum value of a constant C such that conditioning on

$$H(S|X_a, X_b) = H(S|X_b, X_c) = H(S|X_c, X_d) = 0, \tag{4}$$

and

$$H(S|X_a, X_c) = H(S|X_a, X_d) = H(S|X_b) = H(S), \tag{5}$$

we always have

$$H(X_b, X_c) \geq CH(S).$$

We could in principle determine the maximum value of C if Γ_5^* were known, since 5 random variables are involved in the problem. Nevertheless, we can explore the possible values of C by ITIP. By doing so, very quickly we find that ITIP returns a 'True' for all $C \leq 3$, and returns the clause "Not provable by ITIP" for any C slightly greater than 3, say 3.0001. This means that the maximum value of C is lower bounded by 3. This bound is in fact tight, as we now show by the following construction due to Stinson [15]. Let $S_1, S_2, Y_1, Y_2, Y_3, Y_4$, and Y_5 be mutually independent random variables which distribute uniformly on $GF(5)$, and define

$$\begin{aligned} S &= (S_1, S_2) \\ X_a &= (Y_1, Y_4, Y_5 + S_1 + 3S_2) \\ X_b &= (Y_2, Y_5, Y_1 + S_1) \\ X_c &= (Y_3, Y_1, Y_2 + S_2) \\ X_d &= (Y_4, Y_2, Y_3 + S_1 + S_2). \end{aligned}$$

Then it is easy to check that the conditions in both (4) and (5) are satisfied, and that $H(X_b, X_c) = 3H(S)$. Thus the lower bound obtained above is tight.

Using our approach, all information-theoretic bounds reported in the literature for this class of problems can readily be obtained provided that only a definite number of random variables are involved.

All the above examples show how the tools we have discussed in this paper can provide us with tremendous insight into information theory problems. Compared with an information diagram, ITIP is more powerful in the sense that anything one can prove by using an information diagram can as well be proved by ITIP. Moreover, it is also not very practical to use an information diagram for more than 4 random variables[5] unless the random variables form a Markov chain. In a way, ITIP is a "calculator" for information theorists!

However, the advantage of using an information diagram is that one not only can prove a result by using it, but one can very often "see" the result directly from it. Therefore, information diagrams should be used whenever it is possible. Applications of information diagrams to solving more complicated information theory problems can be found in [19][24][28].

From the above examples, we see that many problems in information theory can be reduced to proving certain information inequalities given that the set of random variables involved satisfy an appropriate set of constraints. This point of view does provide a unified way to understand information theory problems of such type. We also see from these examples how the "laws of information theory" govern the impossibilities in information theory. Therefore, the task of searching for new laws is of fundamental importance. Lastly, we see that quantities like $I(X;Y;Z)$, though can take negative values, are indeed integral parts of Shannon's information measures.

6 Entropy and Groups

In this section, we switch gear to discuss an intriguing relation between information theory and group theory [26]. These two seemingly unrelated fields turn out to be intimately related to each other. To avoid heavy notations, we explain this relation by an example with two random variables X_1 and X_2. Consider the inequality

$$H(X_1) + H(X_2) \geq H(X_1, X_2) \tag{6}$$

[5] See [27] for the use of information diagrams for 4 random variables in 2 dimensions.

which involves the (joint) entropies of subsets of X_1 and X_2. If this inequality is valid (which happens to be a basic inequality), then the result in [26] implies that the following inequality is also true for any finite group G and subgroups G_1 and G_2:

$$\log \frac{|G|}{|G_1|} + \log \frac{|G|}{|G_2|} \geq \log \frac{|G|}{|G_1 \cap G_2|} \tag{7}$$

or

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

where we use $|G|$ to denote the order of a group G. The result in [26] also applies in the opposite direction, i.e., if (7) is valid for any G, G_1 and G_2, then (6) is valid for any X_1 and X_2. Using this result, the non-Shannon-type inequality in (1) implies

$$\begin{array}{rcl} |G_3 \cap G_4|^6 |G_1 \cap G_3|^4 & & |G_1||G_2||G_3|^4 |G_4|^4 \\ |G_1 \cap G_4|^4 |G_2 \cap G_3|^4 & \leq & |G_1 \cap G_2|^2 |G_1 \cap G_3 \cap G_4|^5 \\ |G_2 \cap G_4|^4 & & |G_2 \cap G_3 \cap G_4|^5 \end{array}$$

for any finite group G and subgroups G_1, G_2, G_3, and G_4. This inequality appears to be new in group theory. Its meaning is yet to be understood.

7 Concluding Remarks

Z. Zhang and I were not aware of Pippenger's problem when we wrote the series of papers [20][21][22]. After [22] was published, Pippenger read the paper and informed me of his work[6]. On the one hand, Pippenger's work was well ahead of its time. On the other hand, this problem may be one of the few fundamental problems in information theory that Shannon himself did not investigate. Judging from his work in [5][6][7], Han may also have thought about this problem.

As there exist non-Shannon-type information inequalities involving as few as four random variables, it is apparent that our knowledge about the region Γ_n^* is very far from complete. While further characterization of Γ_n^* is of fundamental importance, giving physical interpretations of those already discovered non-Shannon-type inequalities is of equal importance. So far, there has not been too much success along this line except for the work in

[6] Pippenger sent the author the abstract of his talk at the 1986 SPOC Conference which was not included in the proceedings. The reader can contact the author at whyeung@ie.cuhk.edu.hk for a copy of the abstract.

[25]. Nevertheless, the regions Γ_n^* and Γ_n already have found applications in characterizing the coding rate region for multi-source network coding problems in [24] and [27] (Ch. 15).

In probability theory, a fundamental question is, for a set of random variables, what conditional independence relations can be derived from a given set of conditional independence relations, or more generally, what conditional independence relations can co-exist. This is the so-called *probabilistic representability* problem which has been studied extensively [16] [17] [18] [23]. The problem is known to be exceedingly hard, and so far it can be solved only up to 4 random variables [23]. It was shown in [20] that this problem is actually a sub-problem of the characterization of Γ_n^* (so the latter problem is even harder!).

Recently, a one-to-one correspondence between unconstrained information inequalities and a certain type of inequalities governing the order of finite groups and their subgroups has been obtained [26], establishing a strong connection between information theory and group theory. In particular, a new group-theoretic inequality has been derived from the inequality (1).

On the one hand, many important research results have been obtained during the last decade in the study of entropy functions and information inequalities. On the other hand, these results simply reveal how little we have known about the subject. In addition to probability theory and group theory, further research on this subject may also lead to significant contribution to statistical mechanics, where the concept of entropy originated. Quite possibly, we are only at the beginning of yet another adventure in information theory.

A The proof for the equivalence of the Polymatroidal Axioms and the Basic Inequalities

We first show that the polymatroidal axioms imply the basic inequalities. Obviously, (P1) and (P2) imply all entropies are nonnegative. For (P2), by letting $\gamma = \beta\backslash\alpha$, we have $H_\alpha \leq H_{\alpha\cup\gamma}$, or $H(X_\gamma|X_\alpha) \geq 0$. Here, γ and α are non-overlapping subsets of $\mathcal{N}$. For (P3), by letting $\gamma = \beta\backslash\alpha$, $\delta = \alpha \cap \beta$, and $\sigma = \alpha\backslash\beta$, we have $H_{\sigma\cup\delta} + H_{\gamma\cup\delta} \geq H_\delta + H_{\sigma\cup\delta\cup\gamma}$, or $I(X_\sigma; X_\gamma|X_\delta) \geq 0$. Again, σ, δ, and γ are non-overlapping subsets of $\mathcal{N}$. When $\delta = \phi$, from (P3), we have $I(X_\sigma; X_\gamma) \geq 0$. Thus, (P1)-(P3) imply that all entropies are nonnegative, and that all conditional entropies, mutual informations, and conditional mutual informations are nonnegative

provided that the subsets of random variables involved do not overlap. However, for any conditional entropy, mutual information, or conditional mutual information, even if the subsets of random variables involved are overlapping, it can always be written as a linear combination with nonnegative coefficients of entropies, conditional entropies, mutual informations, and conditional mutual informations for which the subsets of random variables involved in any of the latter three types of information measures do not overlap. For example, $I(X_1, X_2; X_1, X_3, X_5|X_3, X_4)$ can be written as $H(X_1|X_3, X_4) + I(X_1, X_2; X_5|X_1, X_3, X_4)$. This shows that (P1)-(P3) imply the basic inequalities.

The converse is trivial and its proof is omitted.

Acknowledgment

I would like to thank Chunxuan Ye for preparing the figures, and Wai Yin Ng, Ueli Maurer, Neil Sloane, and Ning Cai for their useful inputs.

References

[1] C. E. Shannon, "Communication theory of Secrecy systems," *Bell System Technical Journal,* vol. 28, pp. 656-715, Oct 1949.

[2] F. M. Reza, *An Introduction to Information Theory,* MaGraw-Hill, 1961.

[3] Guo-ding Hu, "On the amount of Information," *Teor. Veroyatnost. i Primenen.* vol. 4, pp. 447-455, 1962 (in Russian).

[4] N. M. Abramson, *Information Theory and Coding,* MaGraw-Hill, 1963.

[5] T. S. Han, "Linear dependence structure of the entropy space," *Inform. Contr*, vol. 29, pp. 337-368, 1975.

[6] T. S. Han, "Nonnegative entropy measures of multivariate symmetric correlations," *Inform. Contr.*, vol. 36, pp. 133-156, 1978.

[7] T. S. Han, "A uniqueness of Shannon's information distance and related nonnegativity problems," *Journal of Combinatorics, Information & System Sciences,* vol. 6, no. 4, pp. 320-331, 1981.

[8] I. Csiszár and J. Körner, "Information Theory: Coding Theorem for Discrete Memoryless Systems," New York: Academic Press, and Budapest: Akademiai Kiado, 1981.

[9] A. Papoulis, *Probability, Random Variables and Stochastic Processes,* 2nd ed., McGraw-Hill, 1984.

[10] N. Pippenger, "What are the laws of information theory?" 1986 Special Problems on Communication and Computation Conference, Palo Alto, California, Sept. 3-5, 1986.

[11] T. M. Cover and J. A. Thomas, *Elements of Information Theory.* New York: Wiley, 1991.

[12] R. W. Yeung, "A new outlook on Shannon's information measures," *IEEE Trans. Inform. Theory*, vol. 37, pp. 466-474, May 1991.

[13] T. Kawabata and R. W. Yeung, "The structure of the I-Measure of a Markov chain," *IEEE Trans. Inform. Theory*, vol. 38, pp. 1146-1149, 1992.

[14] R. M. Capocelli, A. De Santis, L. Gargano and U. Vaccaro, "On the size of shares for secret sharing schemes," *J. Cryptography*, vol. 6, pp. 157-167, 1993.

[15] D. R. Stinson, "New general lower bounds on the information rate of secret sharing schemes," Lecture Notes in Comp. Sci. 740, pp 168-182, 1993.

[16] F. Matúš "Probabilistic conditional independence structures and matroid theory: Background," *Int. J. General Systems*, vol. 22, pp. 185-196, 1994.

[17] F. Matúš and M. Studený, "Conditional independences among four random variables I," *Combinatorics, Probability and Computing,* vol. 4, no. 3, pp. 267-278, 1995.

[18] F. Matúš, "Conditional independences among four random variables II," *Combinatorics, Probability & Computing*, vol. 4, pp. 407-417, 1995.

[19] R. W. Yeung, "Multilevel diversity coding with distortion," *IEEE Trans. Inform. Theory*, vol. IT-41, pp. 412-422, Mar 1995.

[20] R. W. Yeung, "A framework for linear information inequalities," *IEEE Trans. Inform. Theory*, vol. 43, pp. 1924-1934, Nov 1997.

[21] Z. Zhang and R. W. Yeung, "A non-Shannon-type conditional information inequality," *IEEE Trans. Inform. Theory*, vol. 43, pp. 1982-1986, Nov 1997.

[22] Z. Zhang and R. W. Yeung, "On Characterization of entropy function via information inequalities," *IEEE Trans. Inform. Theory*, vol. 44, pp. 1440-1452, Jul 1998.

[23] F. Matúš, "Conditional independences among four random variables III: Final conclusion," *Combinatorics, Probability & Computing*, vol. 8, pp. 269-276, 1999.

[24] R. W. Yeung and Z. Zhang, "Distributed source coding for satellite communications," *IEEE Trans. Inform. Theory*, vol. 45, pp. 1111-1120, May 1999.

[25] R. W. Yeung and Z. Zhang, "A class of non-Shannon-type information inequalities and their applications," *Communications in Information and Systems*, vol. 1, no. 1, pp. 87-100, 2001 (http://www.ims.cuhk.edu.hk/čis).

[26] T. H. Chan and R. W. Yeung, "On a relation between information inequalities and group theory," to appear in *IEEE Trans. Inform. Theory.*

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

[28] F.-W.Fu and R. W. Yeung, "On the rate-distortion region for multiple descriptions," to appear in *IEEE Trans. Inform. Theory.*

[29] R. W. Yeung, T. T. Lee and Z. Ye, "An information-theoretic characterization of Markov random fields and its applications," to appear in *IEEE Trans. Inform. Theory.*

[30] I. Sason, "Identification of new enormous classes of non-Shannon type constrained information inequalities and related inequalities for finite groups," submitted to 2002 IEEE International Symposium on Information Theory.

[31] K. Makarychev, Y. Makarychev, A. Romashchenko, and N. Vereshchagin, "A new class of non-Shannon-type inequalities for entropies," submitted to *Communications in Information and Systems.*

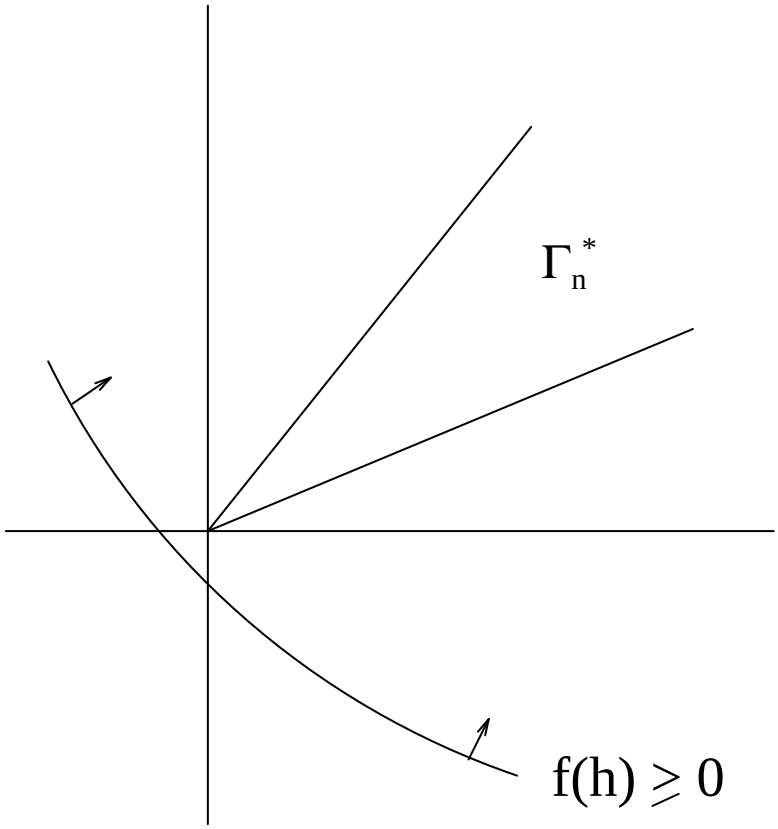

Figure 1: $f(h) \geq 0$ always holds.

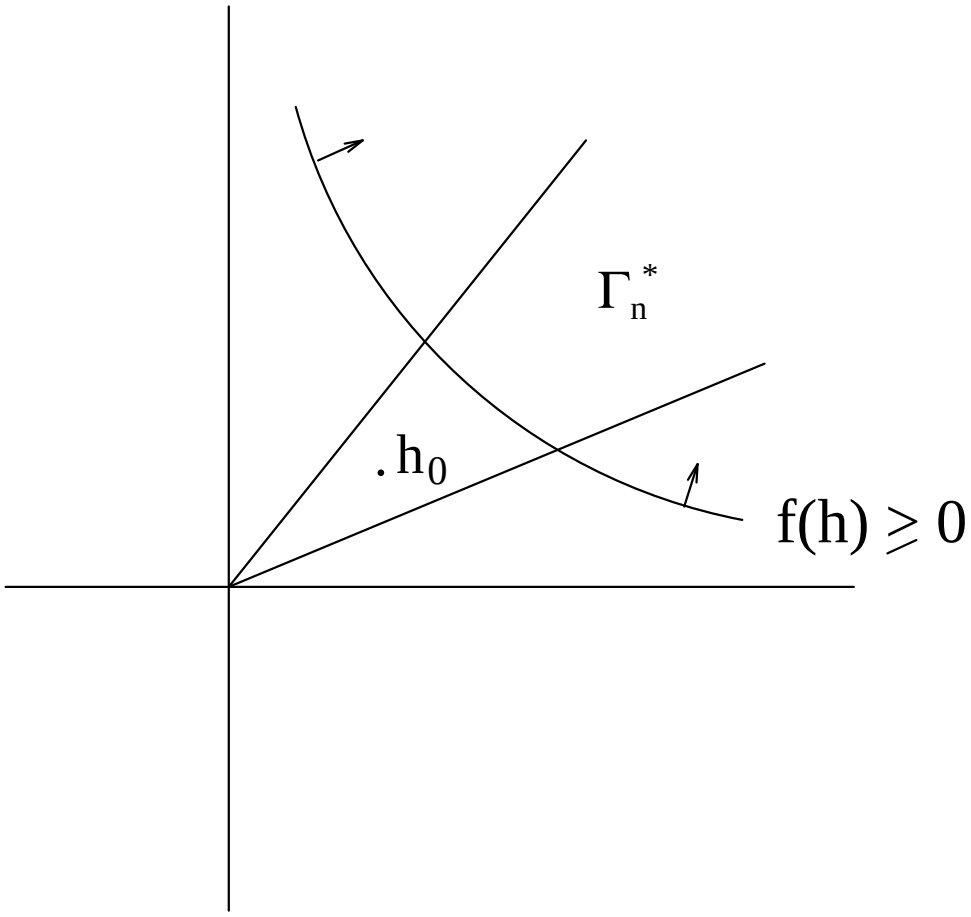

Figure 2: $f(h) \geq 0$ does not always hold.

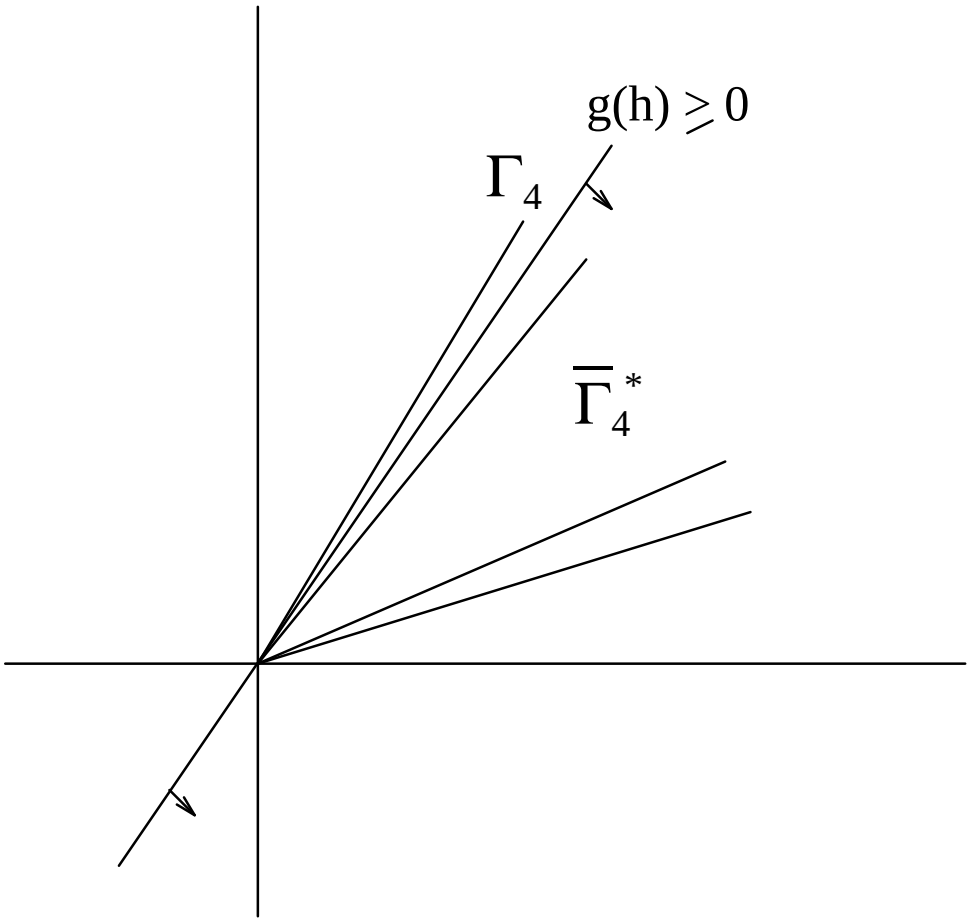

Figure 3: An illustration of $\overline{\Gamma}_4^*$, Γ_4, and $g(h) \geq 0$.

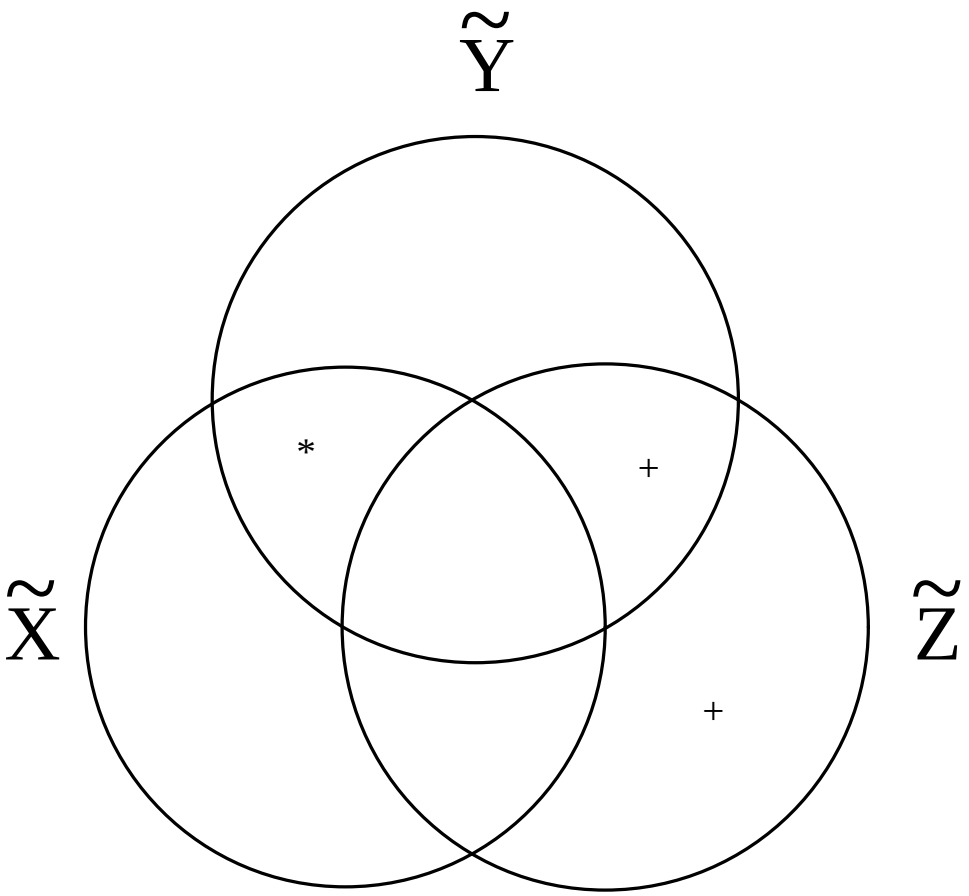

Figure 4: An information diagram for X, Y, and Z.

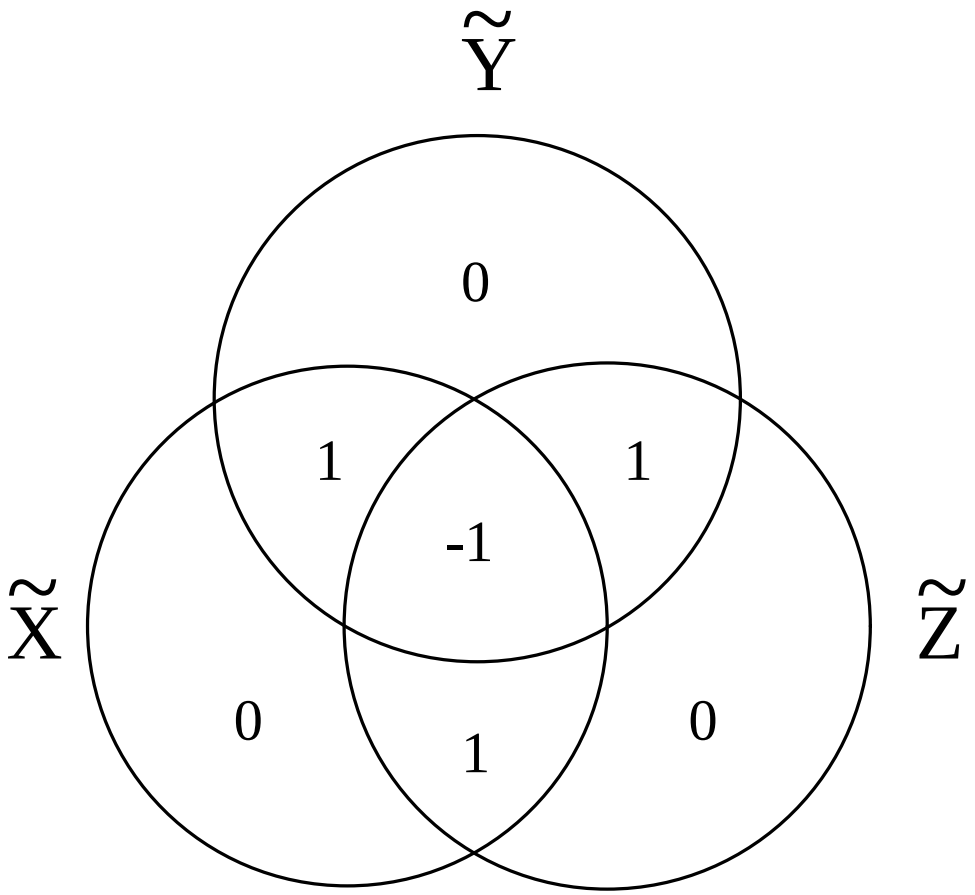

Figure 5: An example for which $I(X;Y;Z)$ is negative.

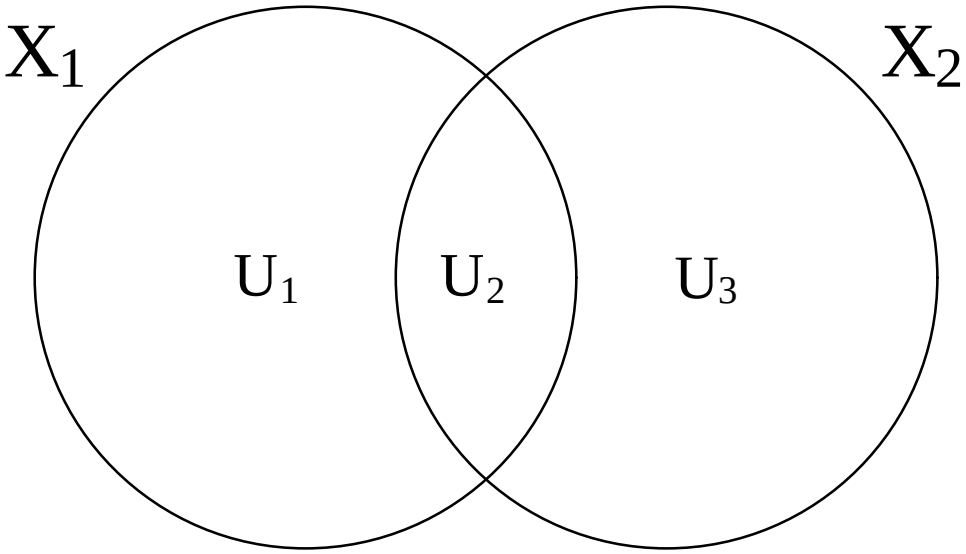

Figure 6: An illustration for the construction of X_1 and X_2

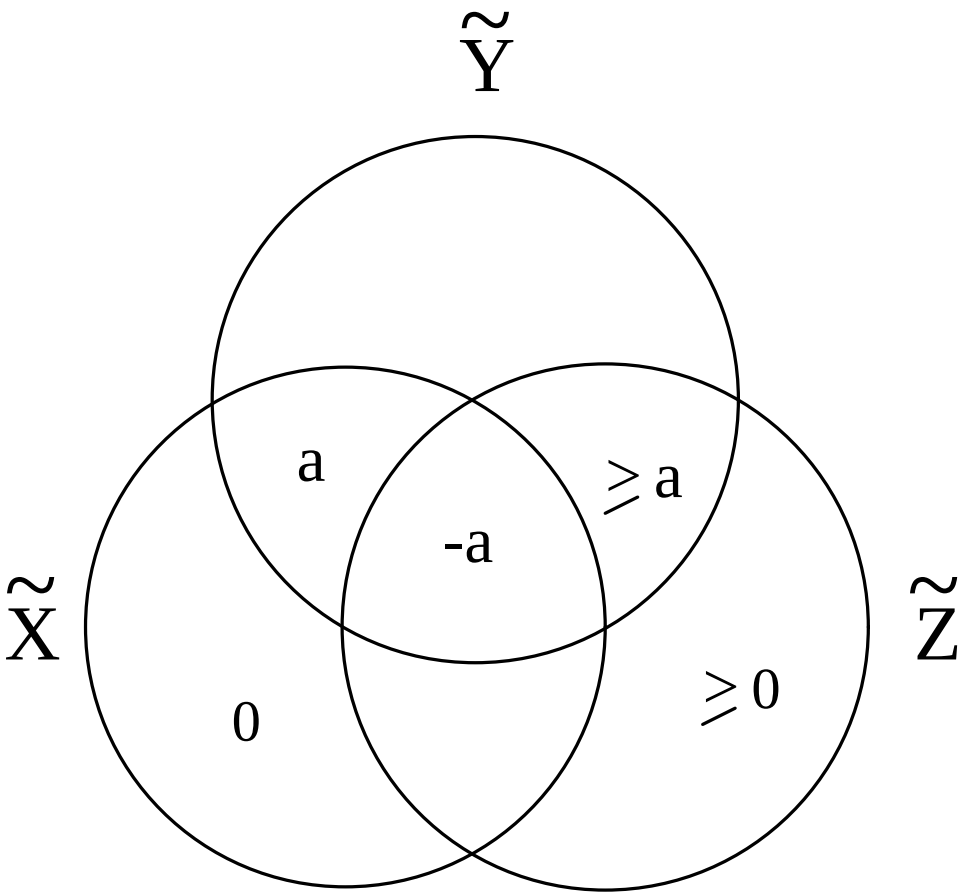

Figure 7: The information diagram for Example 1.

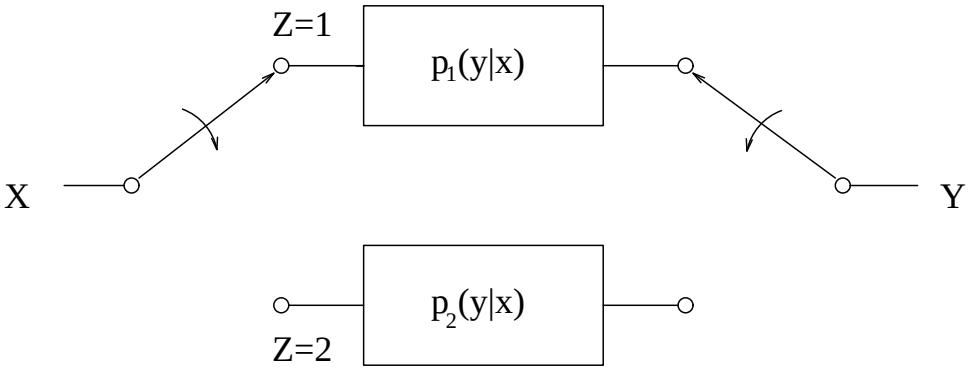

Figure 8: The schematic diagram for Example 2.

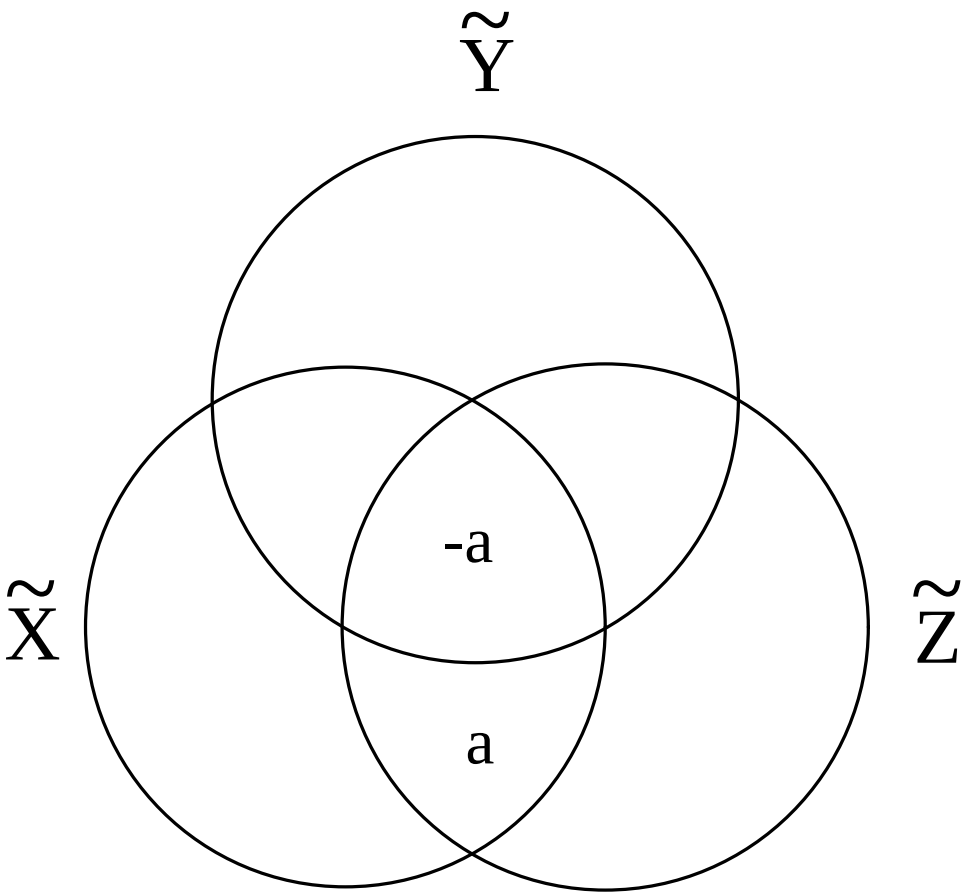

Figure 9: The information diagram for Example 2.

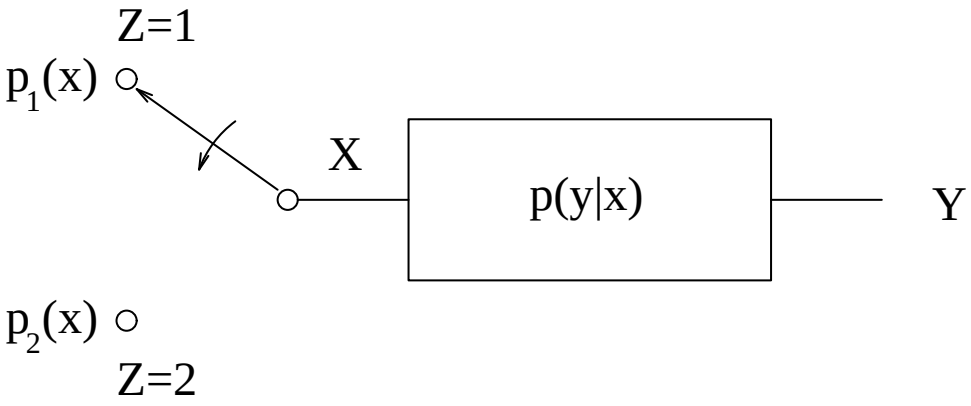

Figure 10: The schematic diagram for Example 3.

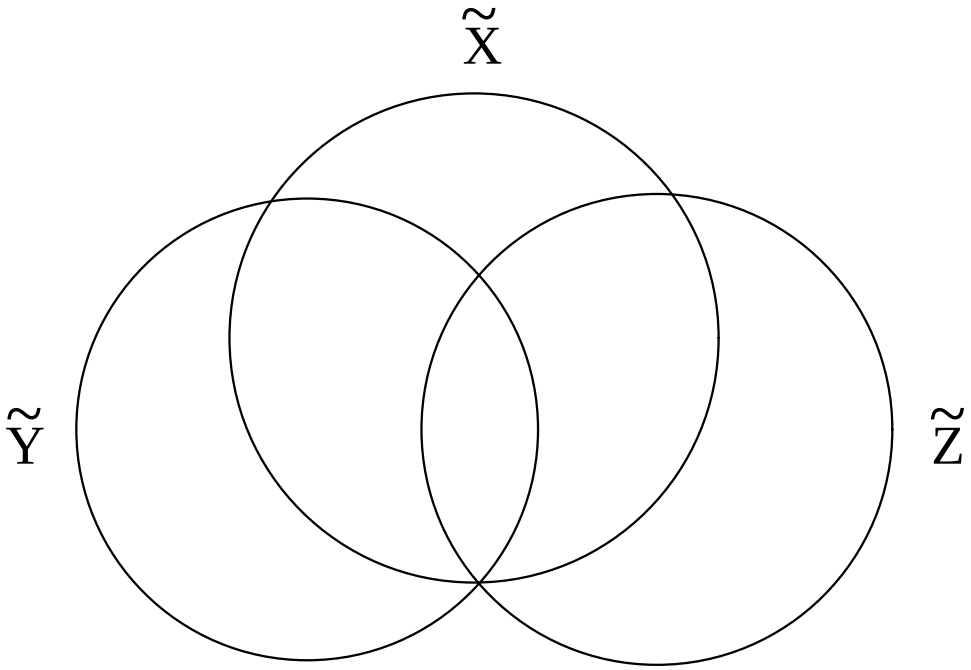

Figure 11: The information diagram for Example 3.

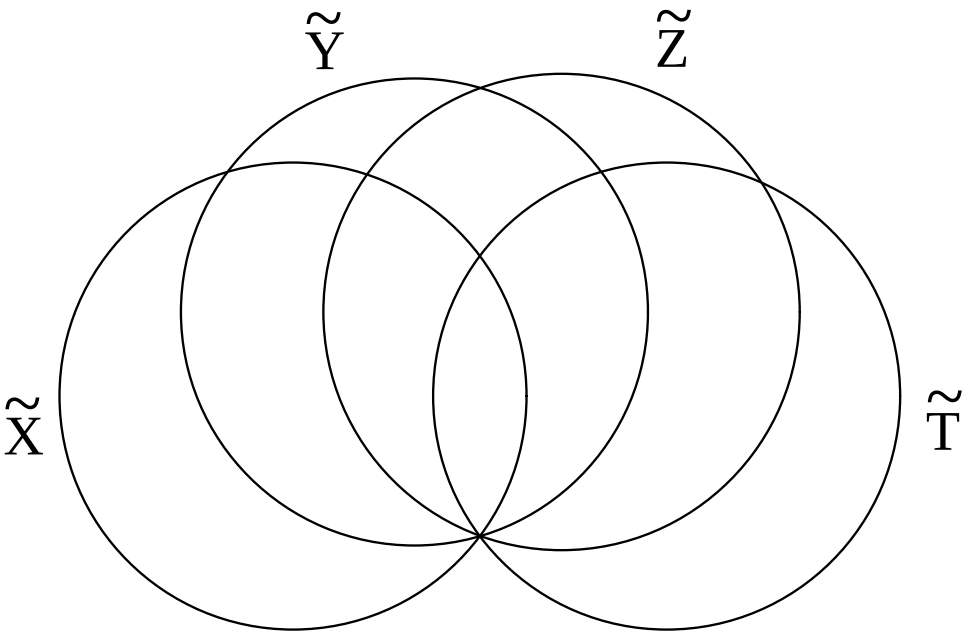

Figure 12: The information diagram for the Markov chain $X \to Y \to Z \to T$.

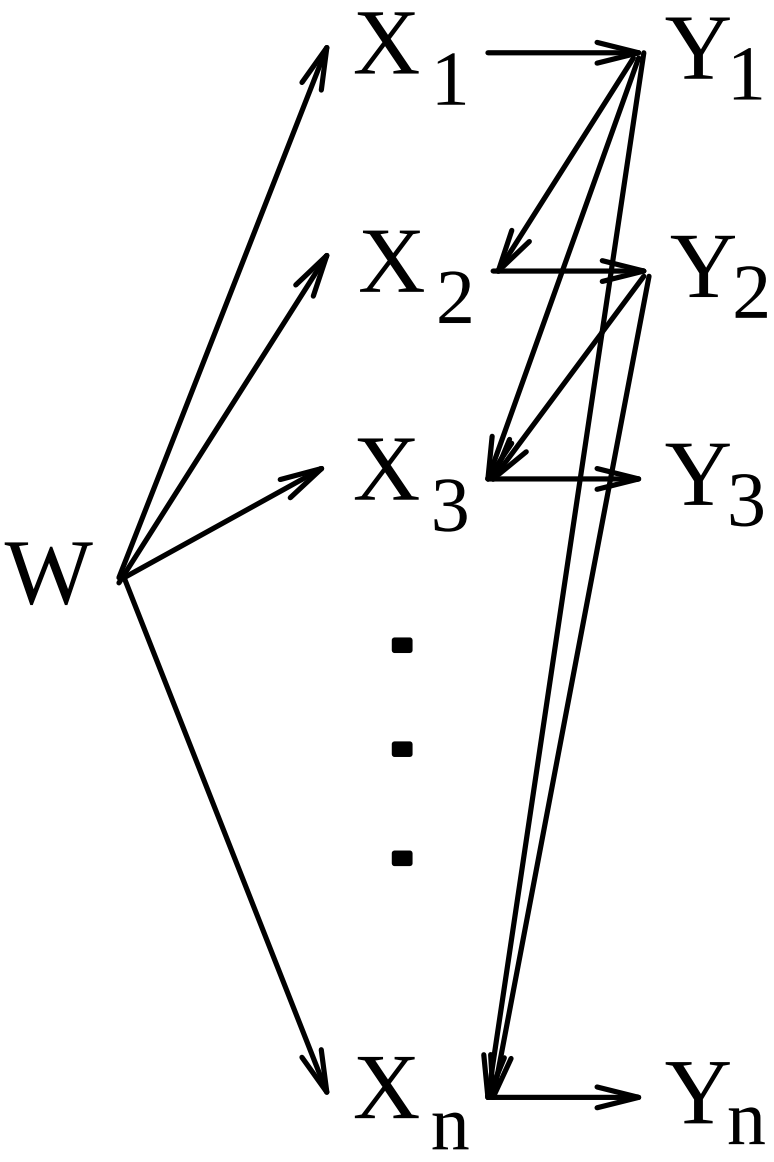

Figure 13: The dependency structure of the random variables involved in the feedback channel problem.