# Proving and Disproving Information Inequalities: Theory and Scalable Algorithms

Siu-Wai Ho, *Senior Member, IEEE*, Lin Ling, Chee Wei Tan, *Senior Member, IEEE*, and Raymond W. Yeung, *Fellow, IEEE*

***Abstract*— Proving or disproving an information inequality is a crucial step in establishing the converse results in coding theorems. However, an information inequality involving more than a few random variables is difficult to be proved or disproved manually. In 1997, Yeung developed a framework that uses linear programming for verifying linear information inequalities. Under the framework, this paper considers a few other problems that can be solved by using Lagrange duality and convex approximation. We will demonstrate how linear programming can be used to find an analytic proof of an information inequality or an analytic counterexample to disprove it if the inequality is not true in general. The way to automatically find a shortest proof or a smallest counterexample is explored. When a given information inequality cannot be proved, the sufficient conditions for a counterexample to disprove the information inequality are found by linear programming. Lastly, we propose a scalable algorithmic framework based on the alternating direction method of multipliers to accelerate solving a multitude of user-specific problems whose overall computational cost can be amortized with the number of users, and present its publicly-available software implementation for large-scale problems.**

***Index Terms*— Entropy, mutual information, information inequality, automated reasoning by convex optimization.**

## I. Introduction

THE importance of computers as a mathematical tool for automated reasoning cannot be overstated as is evident from computer-assisted proofs to derive theorems (e.g., proving all theorems in Principia Mathematica [4]) or to validate well-known mathematical conjectures such as the four color theorem in graph theory by Heesch–Appel–Haken and the Kepler conjecture in geometry by Hales [5]. It is thus imperative to understand how computers can play a similar role of automated reasoning in the realm of information theory. In this paper, we present a theoretical and algorithmic framework to generate computer-aided proofs for information inequalities. We also demonstrate its applicability to a variety of problems and its software implementation as a proof of concept. This paper presents a step toward a systematic understanding of "automated reasoning by convex optimization", where mathematical optimization techniques are used for proof search in large-scale problems.

In information theory, we may need to prove or disprove different kinds of information inequalities in different problems. For example, information inequalities often play a crucial role in establishing the converse of fundamental capacity theorems or Shannon's perfect secrecy theorem. As such, it is important to verify the correctness of a given information inequality by either providing a rigorous proof or otherwise disprove it by providing a counterexample. However, to manually prove or disprove an information inequality is non-trivial in general especially when it involves more than three random variables.

For example, we may want to check the correctness of the following two inequalities:

$$\begin{aligned} &I(A;B|CD) + I(B;D|AC) \\ &\quad\leq \quad I(A;B|D) + I(B;D|A) + H(A) + I(B;D|C) \end{aligned} \tag{1}$$

and

$$\begin{aligned} &I(A;B|CD) + I(B;D|AC) \\ &\quad\leq \quad I(A;B|D) + I(B;D|A) + H(AB|D). \end{aligned} \tag{2}$$

To prove information inequalities, the author in [6] developed a framework for linear information inequalities and provided a software package known as Information Theoretic Inequality Prover (ITIP) [7]. ITIP and XITIP (similar to ITIP but it uses a C-based linear programming solver instead [8]) are widely used software packages, and more recent implementations include minitip [9] and psitip [10]. These software packages can be used to verify an information inequality on a computer. For example, if we use ITIP to verify (1) and (2), we will get "Not provable by ITIP" and "True", respectively. However, after we know that (2) is true, we still need an analytic proof. An analytic proof is the formal way to verify an information inequality and, more importantly, it also provides us further insights about the inequality of interest. One important insight is about the necessary and sufficient conditions for the equality to hold.

Manuscript received May 16, 2019; revised December 22, 2019; accepted March 4, 2020. Date of publication March 23, 2020; date of current version August 18, 2020. This work was supported by the Hong Kong University Grants Committee under Project 11207615 and Project AoE/E-02/08, and in part by the National Natural Science Foundation of China under Grant 61771018. This article was presented at the 2014 IEEE International Symposium on Information Theory, in part at the 2014 IEEE International Conference on Communication Systems, and in part at the 2019 IEEE International Symposium on Information Theory. *(Corresponding author: Lin Ling.)*

Siu-Wai Ho is with the Teletraffic Research Centre, The University of Adelaide, Adelaide, SA 5005, Australia.

Lin Ling and Chee Wei Tan are with the Department of Computer Science, City University of Hong Kong, Hong Kong (e-mail: linling2-c@my.cityu.edu.hk).

Raymond W. Yeung is with the Institute of Network Coding, The Chinese University of Hong Kong, Hong Kong, and also with the Department of Information Engineering, The Chinese University of Hong Kong, Hong Kong.

Communicated by F. Alajaji, Associate Editor for Shannon Theory.

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIT.2020.2982642

Consider (2) again which can be proved by showing

$$\begin{aligned} &I(A;B|D)+I(B;D|A)+H(AB|D)\\ &\quad -I(A;B|CD)-I(B;D|AC) \qquad (3)\\ &=H(B|A,C,D)+H(A|B,C,D)+I(B;C|A)\\ &\quad +I(A;B|D)+I(A;C|D) \qquad (4)\\ &\geq 0, \end{aligned}$$

where (3) can be easily verified by expressing all the qualities on both sides in terms of joint entropies. Then we can further deduce that the equality in (2) holds if and only if all the qualities on the right side of (4) are equal to 0. In Section II, we will demonstrate how to use a linear program to obtain a proof like the one in given above.

Notice that the proof above is not unique, and in fact for any valid information inequality, there are infinitely many proofs, and many of the longer proofs are unnecessarily complicated and hard to gain insights from. Therefore, it would be of interest if we can construct a *shortest proof*, which is defined as the proof involving the least number of elemental inequalities. This is discussed in details in Sections III and V.

In [11], the open problem of whether exact-repair regeneration codes and functional-repair regeneration codes have different rate regions was solved. At the core of the proof is an information inequality involving 16 random variables that is very hard to prove manually. The author tailor-made a linear program to find this required information inequality for the converse result and to render a proof of it.

This interesting result demonstrates that we may need a machine to find a proof when we are dealing with a large-size problem involving many random variables.[1] The framework developed in [6] was applied in [12], [13] for characterizing the rate region of multi-source network coding.[2] Following this line of thought, [14]–[16] have recently developed algorithms for computing the rate regions of network coding problems and their variations.

Now, when an information inequality cannot be proved by ITIP, there are two possible cases. The first case is that the inequality is indeed true but to verify it is outside the capability of ITIP. The existence of such inequalities were first found in [17], [18] and an infinite number of such inequalities were later reported in [19] (see also [20]–[22]). These inequalities are called non-Shannon-type inequalities, which will be explicitly defined in Section IV. Another case is that the given inequality is in fact not true in general. In other words, there exist counterexamples which can disprove the inequality. It is important to distinguish between these two cases. An analytic counterexample is a formal way to disprove a (generally untrue) information inequality subject to a given set of constraints (information inequalities or equalities), and can also provide us with further insights about when this inequality can be made to hold true by manipulating the given set of information constraints.

The existing computational state-of-the-art only *verifies* instead of providing a mathematical proof, i.e., both ITIP and XITIP are in fact only verifiers instead of provers in contrast to what their names suggest. The situation gets even more complex when we study non-Shannon-type inequalities, which are in general a set of infinite number of inequalities [19] (this belongs to semi-infinite programming in the optimization literature) that make the problem dimension much larger and more coupled to be solved by distributed computation. The recent work in [1], [2], [11] demonstrate that Lagrange duality, when applied to the framework in [6], can provide explicitly a proof or a counterexample, and this procedure can in fact be automated to solve large problems involving many random variables. This leads to our online software service AITIP that we describe in details in this paper.

Here we would like to point out a beautiful analogy between linear programming and the task of proving an information inequality, which is illustrated in Fig. 1. We can see that the *geometric* aspect of an information inequality is represented by the primal problem, while the *algebraic* aspect is represented by the corresponding dual problem, and hence there is also a duality between these two aspects of information inequalities. This key insight motivates the linear programming framework in Section II, as it shows that an analytic proof derived from the algebraic properties can be constructed by solving the dual linear program.[3]

Overall, the contributions of the paper are as follows:

1) We develop the mathematical theories and algorithms for proving information theory inequalities using the linear basic inequalities framework and convex optimization theory to prove Shannon's information inequalities.
2) When the given inequality cannot be proved, we show how to find *a smallest counterexample* of (generally untrue) information inequalities by using linear programming to obtain hints for constructing counterexamples involving as few elemental information inequalities as possible.
3) We propose a specialized iterative algorithm based on the alternating direction method of multipliers (ADMM) to make proving and disproving inequalities involving a large number of random variables possible. The algorithm is efficient and parallelizable, and it is designed in a way that part of the computation can be done a priori and cached, which makes it natural to be used in the cloud-computing setting.
4) Our numerical simulations show that, for both randomly generated inequalities and a practical problem in quantum communications, the proposed algorithm is computationally attractive and possesses fast convergence even for a relatively large number of random variables. Our algorithm, implemented in an online software service called AITIP (`https://aitip.org`), can be readily used by anyone in the world.

This paper is organized as follows. Section II shows how to find a proof of an information inequality through

[1] Here we point out a similarity in that the computer-aided verification of Kepler's conjecture is based on minimizing a function with 150 variables with linear programming in a large-scale formal proof project that checks every logical inference of the proof of the Kepler conjecture by computer calculations [5].

[2] The model studied in [12] is a special case of multi-source network coding.

[3] A similar analogy can also be found in [23], where the authors used semidefinite programming to generate certificates of the nonnegativity of polynomials.

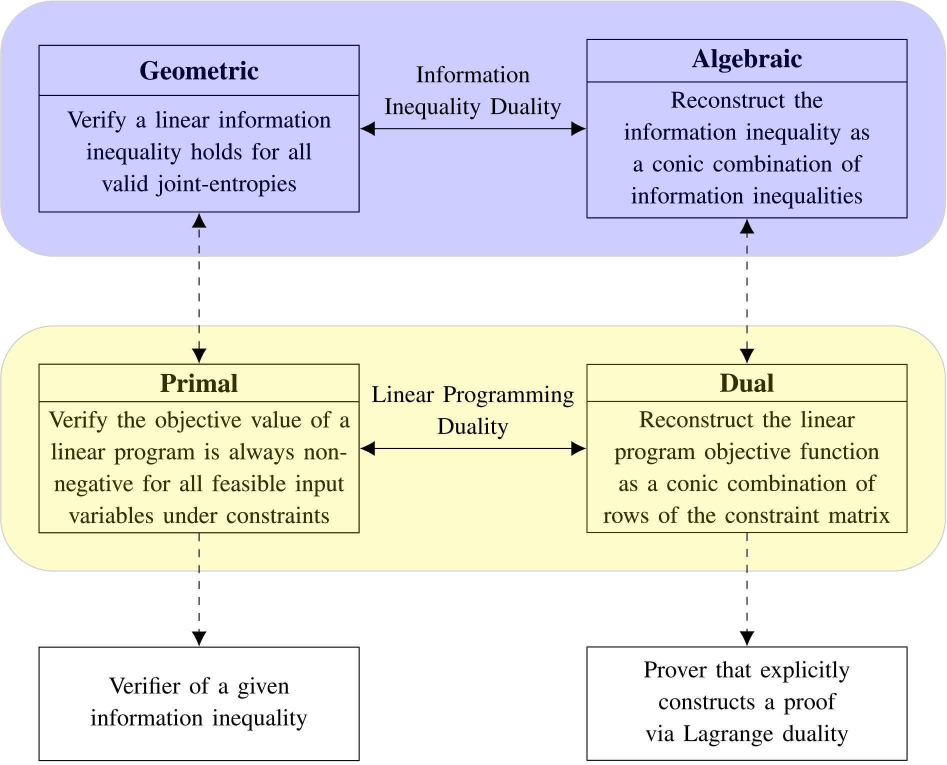

Geometric
Verify a linear information inequality holds for all valid joint-entropies
Information Inequality Duality
Algebraic
Reconstruct the information inequality as a conic combination of information inequalities
Primal
Verify the objective value of a linear program is always non-negative for all feasible input variables under constraints
Linear Programming Duality
Dual
Reconstruct the linear program objective function as a conic combination of rows of the constraint matrix
Verifier of a given information inequality
Prover that explicitly constructs a proof via Lagrange duality

Fig. 1. Block diagram showing the duality between the algebraic and geometric aspects of information inequalities, as well as the correspondence between the information inequalities and their corresponding linear programs. By solving the primal problem, we can verify a given information inequality, and by solving the dual problem, an analytic proof or counterexample can be constructed.

linear programming. In Section III, we introduce the problem formulation of finding the shortest proofs of information inequalities, and propose a convex approximation algorithm that is motivated by the sparse recovery technique in the compressed sensing literature that yields a feasible proof. We illustrate how Lagrange duality can help us to disprove an information inequality in Section IV. In Section V, we provide sufficient conditions for constructing a valid information inequality. In Section VI, we describe a computational framework combining both ADMM and the simplex method that makes proving large scale inequalities possible. Finally, we benchmark the implementation of our proposed algorithm using randomly generated information inequalities as well as a practical problem in quantum communications in Section VII, and give concluding remarks in Sections VIII and IX.

## II. Linear Information Inequalities Framework

In the current and the next section, we introduce the framework for constructing proofs and, in particular, the shortest proofs using linear programming. A high-level overview of the framework is shown in Fig. 2. The frameworks in Sections IV and V for constructing counterexamples and the smallest counterexamples are very similar.

Consider $n$ random variables $(X_1, X_2, \ldots, X_n)$ and the (joint) entropies of all the non-empty subset of these random variables form a column vector $\mathbf{h}$. For example, if $n = 3$, then

$$\begin{aligned}\mathbf{h} = (&H(X_1), H(X_2), H(X_3), H(X_1, X_2), H(X_2, X_3),\\ &H(X_1, X_3), H(X_1, X_2, X_3)). \qquad (5)\end{aligned}$$

The coefficients related to an information inequality can be denoted by a column vector $\mathbf{b}$. To illustrate, continue the example of $\mathbf{h}$ in (5). Then, the information inequality $-H(X_1, X_3) + H(X_1, X_2, X_3) \geq 0$ is denoted by

$$\mathbf{b}^T\mathbf{h} \geq 0$$

with $\mathbf{b} = [0\ 0\ 0\ 0\ 0\ -1\ 1]^T$. Due to the nonnegativity of Shannon's information measures, we know that $\mathbf{h}$ must satisfy certain inequalities. For example,

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

The set of all the inequalities due to the nonnegativity of Shannon's information measures is defined as the *basic inequalities*. Note that this set is not minimal in the sense that some basic inequalities can be implied by others. Let $\mathcal{N} = \{1, 2, \ldots, n\}$, a minimal subset of the basic inequalities is defined as the *elemental inequalities* [24, P. 340], namely

$$H(X_i|X_{\mathcal{N}-i}) \geq 0 \qquad (6)$$

$$I(X_i; X_j|X_{\mathcal{K}}) \geq 0 \text{ where } i \neq j \text{ and } \mathcal{K} \subseteq \mathcal{N} - \{i, j\} \qquad (7)$$

and these elemental inequalities are denoted by

$$\mathbf{Dh} \geq \mathbf{0} \qquad (8)$$

in this paper. Obviously, any vector $\mathbf{h}$ must satisfy (8). An important property about this set is that all the inequalities due to the nonnegativity of Shannon's information measures, like $H(X_1) \geq 0$, $H(X_1, X_2|X_3) \geq 0$, etc., can be obtained as a conic combination (also known as a nonnegative linear combination) of the elemental inequalities. Therefore, an information inequality can be proved by using the nonnegativity of Shannon's information measures if and only if the inequality

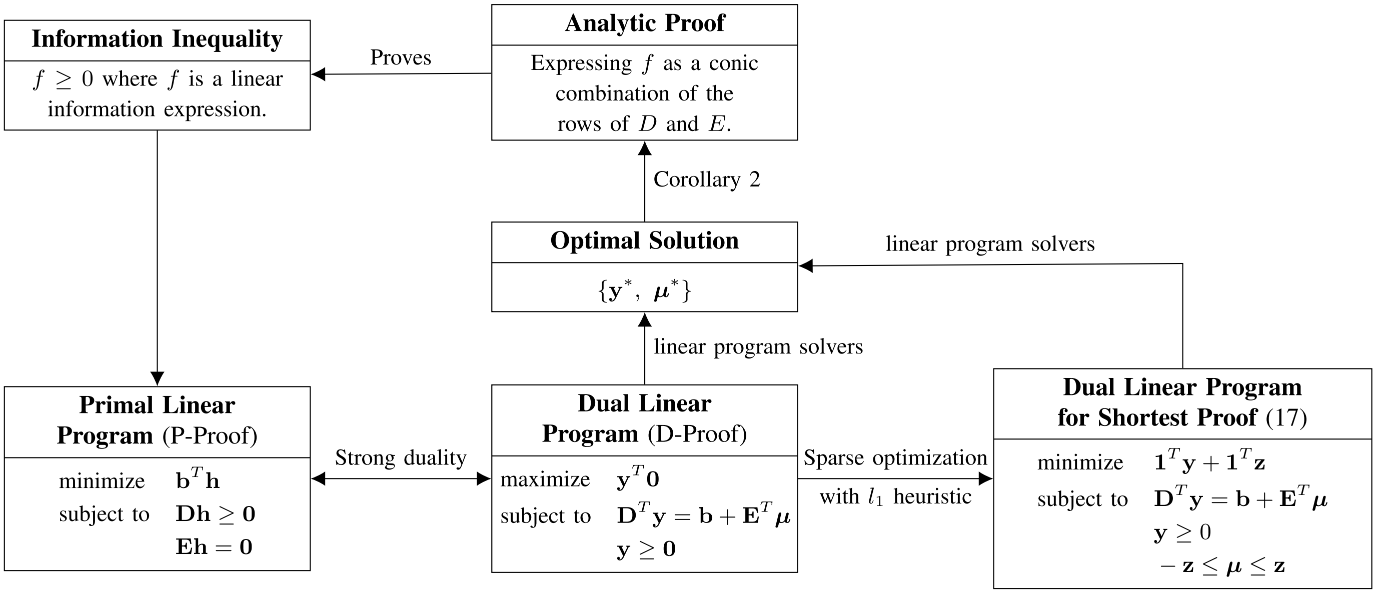

Information Inequality
$f \geq 0$ where $f$ is a linear information expression.
Proves
Analytic Proof
Expressing $f$ as a conic combination of the rows of $D$ and $E$.
Corollary 2
Optimal Solution
$\{\mathbf{y}^*,\ \boldsymbol{\mu}^*\}$
linear program solvers
linear program solvers
Primal Linear Program (P-Proof)
minimize $\mathbf{b}^T\mathbf{h}$
subject to $\mathbf{D}\mathbf{h} \geq \mathbf{0}$
$\mathbf{E}\mathbf{h} = \mathbf{0}$
Strong duality
Dual Linear Program (D-Proof)
maximize $\mathbf{y}^T\mathbf{0}$
subject to $\mathbf{D}^T\mathbf{y} = \mathbf{b} + \mathbf{E}^T\boldsymbol{\mu}$
$\mathbf{y} \geq \mathbf{0}$
Sparse optimization
with $l_1$ heuristic
Dual Linear Program for Shortest Proof (17)
minimize $\mathbf{1}^T\mathbf{y} + \mathbf{1}^T\mathbf{z}$
subject to $\mathbf{D}^T\mathbf{y} = \mathbf{b} + \mathbf{E}^T\boldsymbol{\mu}$
$\mathbf{y} \geq 0$
$-\mathbf{z} \leq \boldsymbol{\mu} \leq \mathbf{z}$

Fig. 2. Block diagram illustrating how an analytic proof and its smallest possible version can be constructed using linear programming. The process of constructing counterexamples and the smallest counterexample is very similar. Example 1 illustrates how the proof to the classical perfect secrecy problem can be automated.

can be implied by the elemental inequalities. An information inequality, which is implied by the nonnegativity of Shannon's information measures, is called *Shannon-type inequality*.

Very often we want to prove an information inequality $\mathbf{b}^T\mathbf{h} \geq 0$ subject to a given set of equality constraints $\mathbf{E}\mathbf{h} = \mathbf{0}$. When there is no equality constraint, this set is simply empty.

The linear combination of the joint entropies $\mathbf{b}^T\mathbf{h}$ is a valid information inequality if and only if it is always nonnegative [6]. Consider the following linear program:

$$\begin{array}{ll} \text{minimize} & \mathbf{b}^T\mathbf{h} \\ \text{subject to} & \mathbf{D}\mathbf{h} \geq \mathbf{0} \\ & \mathbf{E}\mathbf{h} = \mathbf{0} \\ \text{variables:} & \mathbf{h}, \end{array} \qquad \text{(P-Proof)}$$

and its Lagrange dual problem (also a linear program):

$$\begin{array}{ll} \text{maximize} & \mathbf{y}^T\mathbf{0} \\ \text{subject to} & \mathbf{D}^T\mathbf{y} = \mathbf{b} + \mathbf{E}^T\boldsymbol{\mu} \\ & \mathbf{y} \geq \mathbf{0} \\ \text{variables:} & \mathbf{y},\ \boldsymbol{\mu}. \end{array} \qquad \text{(D-Proof)}$$

In Section I, we claim that the primal problem (P-Proof) and dual problem (D-Proof) represent the geometric and algebraic aspects of the problem of proving information inequalities, respectively. Now with the two problem formally formulated, we can explain the connections in more details. The optimal value of (P-Proof) is zero if $\mathbf{b}^T\mathbf{h} \geq 0$ is a Shannon-type inequality, and is $-\infty$ otherwise [24, Theorem 14.4]. The two cases are illustrated geometrically in Fig. 3, where $\Gamma$ represents the feasible region of (P-Proof). Because we are comparing the optimal value of a linear program with 0, the primal problem represents the geometric aspect. The dual constraint in (D-Proof), $\mathbf{D}^T\mathbf{y} = \mathbf{b} + \mathbf{E}^T\boldsymbol{\mu}$, can be interpreted as reconstructing $\mathbf{b}$ as a linear combination of rows in constraint matrices $\mathbf{D}$ and $-\mathbf{E}$ by finding the weights of the rows, and therefore the dual problem is considered the algebraic aspect.

From the strong duality of linear programming, i.e., the duality gap is zero in linear programs [25, Theorem 4.4], the optimal values of (P-Proof) and (D-Proof) are equal. Furthermore, we have the following optimality results which is used to show an analytic proof for any Shannon-type inequality.[4]

*Theorem 1:* The inequality $\mathbf{b}^T\mathbf{h} \geq 0$ is a Shannon-type inequality if and only if the feasibility problem (D-Proof) is feasible, i.e., there exists a feasible solution $[\mathbf{y}^{*T}\ \boldsymbol{\mu}^{*T}]^T$ to (D-Proof) satisfying

$$\mathbf{D}^T\mathbf{y}^* = \mathbf{b} + \mathbf{E}^T\boldsymbol{\mu}^* \text{ and } \mathbf{y}^* \geq \mathbf{0}. \tag{9}$$

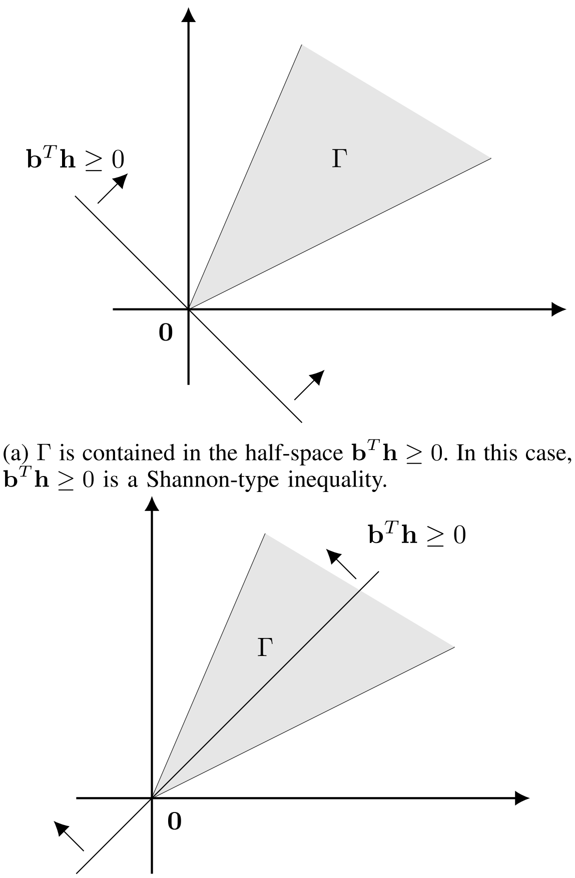
*Proof:* If (9) is true, then $\mathbf{b}^T\mathbf{h} = (\mathbf{y}^*)^T\mathbf{D}\mathbf{h} - (\boldsymbol{\mu}^*)^T\mathbf{E}\mathbf{h}$. Hence, $\mathbf{b}^T\mathbf{h} \geq 0$ follows from that $\mathbf{D}\mathbf{h} \geq \mathbf{0}$, $\mathbf{E}\mathbf{h} = \mathbf{0}$ and $\mathbf{y}^* \geq 0$.

If $\mathbf{b}^T\mathbf{h} \geq 0$ is a Shannon-type inequality, then the optimization problem in (P-Proof) has an optimal value equal to zero [24, Theorem 14.4]. Using the Karush-Kuhn-Tucker (KKT) conditions, the respective optimal primal and dual solutions $\mathbf{h}^*$ and $[\mathbf{y}^{*T}\ \boldsymbol{\mu}^{*T}]^T$ satisfy

$$\begin{aligned} \mathbf{D}^T\mathbf{y}^* = \mathbf{b} + \mathbf{E}^T\boldsymbol{\mu}^*, &\quad \text{(Stationarity of Lagrangian)} \\ \mathbf{y}^{*T}\mathbf{D}\mathbf{h}^* = \mathbf{0}, &\quad \text{(Complementary slackness)} \\ \mathbf{D}\mathbf{h}^* \geq \mathbf{0}, \mathbf{E}\mathbf{h}^* = \mathbf{0}, \mathbf{h}^* \geq \mathbf{0}, &\quad \text{(Primal feasibility)} \\ \mathbf{D}^T\mathbf{y}^* = \mathbf{b} + \mathbf{E}^T\boldsymbol{\mu}^*, \mathbf{y}^* \geq \mathbf{0}. &\quad \text{(Dual feasibility)} \end{aligned}$$

Hence, (9) follows because it is the stationarity of the Lagrangian in the KKT conditions. It is easy to see that the complementary slackness is implied by the stationarity of Lagrangian, the primal feasibility and the fact that $\mathbf{b}^T\mathbf{h}^* = 0$. From the KKT conditions, we have $\mathbf{b}^T\mathbf{h}^* - \mathbf{y}^{*T}\mathbf{D}\mathbf{h}^* + \boldsymbol{\mu}^{*T}\mathbf{E}\mathbf{h}^* = 0$. But this implies $-\mathbf{y}^{*T}\mathbf{D}\mathbf{h} + \boldsymbol{\mu}^{*T}\mathbf{E}\mathbf{h} \leq 0 \leq$

[4] The idea of using the Lagrange duality to find an analytic proof has also been used in [11].


$\mathbf{b}^T\mathbf{h} \geq 0$
$\Gamma$
$\mathbf{0}$
$\mathbf{b}^T\mathbf{h} \geq 0$
$\Gamma$
$\mathbf{0}$

(a) $\Gamma$ is contained in the half-space $\mathbf{b}^T\mathbf{h} \geq 0$. In this case, $\mathbf{b}^T\mathbf{h} \geq 0$ is a Shannon-type inequality.

(b) $\Gamma$ is not contained in the half-space $\mathbf{b}^T\mathbf{h} \geq 0$. In this case, $\mathbf{b}^T\mathbf{h} \geq 0$ is either not true in general or a non-Shannon-type inequality.

Fig. 3. Geometric illustration of the relation between the feasible region $\Gamma$ and the half-space $\mathbf{b}^T\mathbf{h} \geq 0$, assuming that there is no problem-specific constraint, i.e., $\mathbf{E}$ is empty. The figures considering problem-specific constraints can only be drawn in 3-D and can be found in [24, 13.3.2].

$\mathbf{b}^T\mathbf{h}$ for all feasible $\mathbf{h}$. This proves the last part of the theorem. □

Then we can have the following direct consequence.

*Corollary 1:* If $\mathbf{b}^T\mathbf{h} \geq 0$ is a Shannon-type inequality, then let $[\mathbf{y}^{*T}\ \boldsymbol{\mu}^{*T}]^T$ be the optimal solution to (D-Proof) and an analytic proof can be written as follows. For all feasible $\mathbf{h}$,

$$\begin{aligned}\mathbf{b}^T\mathbf{h} &= (\mathbf{y}^*)^T\mathbf{D}\mathbf{h} - (\boldsymbol{\mu}^*)^T\mathbf{E}\mathbf{h}\\ &\geq 0, \qquad (10)\end{aligned}$$

where (10) follows from that $\mathbf{y}^* \geq 0$, $\mathbf{D}\mathbf{h} \geq \mathbf{0}$ and $\mathbf{E}\mathbf{h} = \mathbf{0}$.

As an illustrative example, let us apply Corollary 1 on Shannon's perfect secrecy theorem using our AITIP software service (cf. Section VI).

*Example 1:* Suppose we want to prove $H(U) \leq H(R)$ subject to $I(U;X) = 0$ and $H(U|R,X) = 0$. AITIP gives the following output:

$AITIP('H(U) <= H(R)',' I(U;X) = 0',' H(U|R,X) = 0')$

True. The inequality follows from

$$\begin{aligned}-H(U) + H(R) &= (-H(U,X) + H(U,R,X)) +\\ &\quad (H(R) + H(X) - H(R,X)) +\\ &\quad \{-H(U) - H(X) + H(U,X)\} +\\ &\quad \{H(R,X) - H(U,R,X)\}\\ &\geq 0,\end{aligned}$$

where $(\cdot)$ is nonnegative as it is conditional entropy, mutual information or conditional mutual information. All $\{\cdot\}$ are equal to 0 due to the given constraints. Equality holds iff all $(\cdot)$ are equal to 0.

## III. The Shortest Proofs

We have seen that an information inequality can be proved by expressing it into a linear combination of elemental inequalities. However, there can be more than one way to express the same information inequality. For example,

$$\begin{aligned}&H(X,Y,Z) - H(X|Y,Z) - H(Y|X,Z) - H(Z|X,Y)\\ &= I(X;Y) + I(X;Z|Y) + I(Y;Z|X) \qquad (11)\\ &= I(X;Z) + I(X;Y|Z) + I(Y;Z|X) \qquad (12)\\ &= I(Y;Z) + I(X;Z|Y) + I(X;Y|Z). \qquad (13)\end{aligned}$$

So we can prove $H(X,Y,Z) - H(X|Y,Z) - H(Y|X,Z) - H(Z|X,Y) \geq 0$ by proving any of the equalities in (11)–(13). In fact,

$$\begin{aligned}&H(X,Y,Z) - H(X|Y,Z) - H(Y|X,Z) - H(Z|X,Y)\\ &= 0.8(I(X;Y) + I(X;Z|Y) + I(Y;Z|X)) +\\ &\quad 0.1(I(X;Z) + I(X;Y|Z) + I(Y;Z|X)) +\\ &\quad 0.1(I(Y;Z) + I(X;Z|Y) + I(X;Y|Z)) \qquad (14)\end{aligned}$$

is also true. Obviously, the proof given in (14) is unnecessarily complicated and long as well as not succinctly elegant enough.

A *shortest proof* of an information inequality is considered as a proof involving the least number of elemental inequalities. For a given Shannon-type information inequality, there often exist multiple shortest proofs. Let us consider the following combinatorial optimization problem:

$$\begin{array}{ll}\text{minimize} & \|[\mathbf{y}^T\ \boldsymbol{\mu}^T]^T\|_0\\ \text{subject to} & \mathbf{D}^T\mathbf{y} = \mathbf{b} + \mathbf{E}^T\boldsymbol{\mu},\ \mathbf{y} \geq \mathbf{0},\\ \text{variables:} & \mathbf{y},\ \boldsymbol{\mu},\end{array} \qquad (15)$$

where $\|\mathbf{x}\|_0$ is the cardinality or the number of nonzero components in the vector $\mathbf{x}$. Now, (15) is a combinatorial problem that is generally hard to solve. Suppose there exists a feasible dual variable $[\mathbf{y}^T\ \boldsymbol{\mu}^T]^T$. Consider the following nonempty and bounded polyhedron:

$$P = \{[\mathbf{y}^T\ \boldsymbol{\mu}^T]^T \in \mathcal{R}^{m+q} \mid \mathbf{y} \geq \mathbf{0},\ \mathbf{D}^T\mathbf{y} = \mathbf{b} + \mathbf{E}^T\boldsymbol{\mu}\},$$

where $m$ and $q$ are the number of rows in $\mathbf{D}$ and $\mathbf{E}$, respectively. From Lemma 2 in [26], (15) has a solution that is a vertex of $P$. So rather than solving (15) directly as an optimization problem, we can also enumerate all the vertices of $P$. Then the vertex that has the least cardinality is guaranteed to be a solution to (15). A practical pivot-based algorithm has been proposed in [27] to find the $v$ vertices of a polyhedron in $\mathbf{R}^d$ defined by a non-degenerate system of $m$ inequalities in $O(mdv)$ time and $O(md)$ space. But vertex enumeration can be computationally inefficient especially in large problems.

The marriage of the information inequalities framework and recent developments in convex approximation for sparse recovery offers new directions to explore. Now, we consider the following convex approximation problem to tackle (15):

$$\begin{array}{rl} \text{minimize} & \mathbf{1}^T\mathbf{y} + \|\boldsymbol{\mu}\|_1 \\ \text{subject to} & \mathbf{D}^T\mathbf{y} = \mathbf{b} + \mathbf{E}^T\boldsymbol{\mu},\ \mathbf{y} \geq \mathbf{0} \\ \text{variables:} & \mathbf{y},\ \boldsymbol{\mu}. \end{array} \tag{16}$$

Here (16) is obtained by approximating the cardinality term $\|\mathbf{x}\|_0$ with its convex envelope, $\|\mathbf{x}\|_1$. This $l_1$-norm heuristic has been used extensively in the compressed sensing area [28], and it is known that under certain conditions this approximation is exact. Unfortunately in our case (15) and (16) are not equivalent, as we will see in Example 2. Since (16) is a (nonsmoothed) convex optimization problem, it can in principle be solved by an interior-point method numerically [29] or the homotopy algorithm in [30]. However, (16) can be further transformed to the following equivalent problem that is a linear program:

$$\begin{array}{rl} \text{minimize} & \mathbf{1}^T\mathbf{y} + \mathbf{1}^T\mathbf{z} \\ \text{subject to} & \mathbf{D}^T\mathbf{y} = \mathbf{b} + \mathbf{E}^T\boldsymbol{\mu},\ \mathbf{y} \geq \mathbf{0} \\ & -\mathbf{z} \leq \boldsymbol{\mu} \leq \mathbf{z} \\ \text{variables:} & \mathbf{y},\ \boldsymbol{\mu},\ \mathbf{z}. \end{array} \tag{17}$$

Note that the convex approximation problem in (16) can be different from the original problem in (15) as shown below.

*Example 2:* Consider

$$H(Y,Z) - H(Y|X,Z) - H(Z|X,Y) + I(X;Y|Z) \tag{18}$$
$$= I(X;Z) + 2I(X;Y|Z) + I(Y;Z|X) \tag{19}$$
$$= I(X;Y) + I(X;Y|Z) + I(Y;Z|X) + I(X;Z|Y). \tag{20}$$

In order to prove the nonnegativity of (18), we just need to show the equality in either (19) or (20). The summation of coefficients in both (19) and (20) are the same and equal to 4. However, (19) involves less number of Shannon's information measures. Therefore, the optimization problems in (15) and (16) can have different results for some inequalities.

Now, we consider a slight modification of the linear program in (17) to obtain further results. Suppose we know that an information inequality is correct if the set of equality constraints $\mathbf{Eh} = \mathbf{0}$ is assumed. It is interesting to know the minimal set of equality constraints that is required. In other words, we want to know the minimal number of rows in $\mathbf{E}$ which are sufficient to prove the same inequality. To solve this problem, we just need to remove $\mathbf{y}^T$ in the objective function in (15). Since this problem is also NP-hard, we will consider the approximation of this problem by replacing $\mathbf{1}^T\mathbf{y}+\mathbf{1}^T\mathbf{z}$ by $\mathbf{1}^T\mathbf{z}$ in the linear program in (17). The following example revisits the implication problem in [24, Ex. 14.9] to demonstrate a much shorter proof. The modified linear program shows that not all the equality constraints in [24, (14.73)] are necessary.

*Example 3:* Under the assumption that

$$\begin{aligned} 0 &= I(X;Y|Z) = I(X;T|Y) = I(X;Z|Y) \\ &= I(X;T|Z) = I(X;Z|T), \end{aligned} \tag{21}$$

we want to prove $I(X;Y|T) = 0$. The modified linear program shows

$$\begin{aligned} -I(X;Y|T) &= I(X;T|Z) + I(X;Z|Y) - \\ &\quad I(X;Y|Z) - I(X;T|Y) - I(X;Z|T) \quad (22) \\ &\geq 0, \quad (23) \end{aligned}$$

where (23) follows from using only $I(X;Y|Z) = I(X;T|Y) = I(X;Z|T) = 0$ in (21). This proof can provide us further insights. By rearranging the terms in (22), $I(X;T|Z) = I(X;Z|Y) = 0$ can be implied by just assuming $I(X;Y|Z) = I(X;T|Y) = I(X;Z|T) = 0$. This further explains why some equality constraints in (21) are redundant.

Before the end of this section, we want to remark that a proof can be shorter if $\mathbf{D}$ is expanded to include all the inequalities due to the nonnegativity of Shannon's information measures. This expansion is equivalent to expanding $\mathbf{D}$ by adding some positive linear combinations of the rows in $\mathbf{D}$. If this new $\mathbf{D}$ is used in the linear program in (P-Proof), the same optimal solution will be obtained and hence, we can still obtain Theorem 1. However, we can now obtain a shorter proof. For example,

$$H(X,Y,Z) - I(Y;Z|X) - H(Z|X,Y) \tag{24}$$
$$= H(X|Y,Z) + H(Y|X,Z) + I(X;Y) + I(X;Z|Y) \tag{25}$$
$$= H(X) + H(Y|X,Z). \tag{26}$$

When we prove the nonnegativity of (24), a longer proof in (25) is obtained if $\mathbf{D}$ contains only elemental inequalities. However, a shorter proof can be shown in (26) if the linear program *learns* more possibilities of conic combination through the matrix $\mathbf{D}$. However, we need to pay the price for a larger size in $\mathbf{D}$ that affects the computational time to solve the linear programs.

## IV. Disproving an Information Inequality

We have seen how to use a linear program to obtain a proof of a Shannon-type inequality. In this section, we explore how to disprove an information inequality. Suppose we are given an information inequality $\mathbf{b}^T\mathbf{h} \geq 0$ which cannot be proved by the linear program (D-Proof). In other words, the linear program does not give zero as the optimal objective value [24, Theorem 14.4]. It is still unclear whether a) $\mathbf{b}^T\mathbf{h} < 0$ for some $\mathbf{h}$ or b) $\mathbf{b}^T\mathbf{h} \geq 0$ is indeed true for all feasible $\mathbf{h}$, but proving it is beyond the capability of the linear program using $\mathbf{D}$ (in the case, the inequality cannot be expressed as $\mathbf{D}^T\mathbf{y}$ for some $\mathbf{y} \geq \mathbf{0}$ and $\mathbf{b}^T\mathbf{h} \geq 0$ is called a *non-Shannon-type inequality* [18]). Therefore, we need a counterexample to explicitly disprove $\mathbf{b}^T\mathbf{h} \geq 0$.

Suppose we can find an $\mathbf{h}$ such that $\mathbf{Dh} \geq \mathbf{0}$ and $\mathbf{b}^T\mathbf{h} < 0$. This is still insufficient to be a counterexample because there may not exist any joint distribution $P_{X_1,X_2,\ldots,X_n}$ that realizes $\mathbf{h}$. An example is shown in [24, (15.85)]. In general, there is no known algorithm to construct $P_{X_1,X_2,\ldots,X_n}$ from any given $\mathbf{h}$, and hence, it seems that finding $\mathbf{h}$ may not give any immediate help. In the following, we will show that a linear program for finding $\mathbf{h}$ is still useful as long as we know how to read the information contained in $\mathbf{h}$. Its dual problem will give us the sufficient conditions for the counterexample to disprove $\mathbf{b}^T\mathbf{h} \geq 0$.

Suppose the last element in $\mathbf{h}$ denotes the joint entropy $H(X_1, X_2, \ldots, X_n)$. Let $\mathbf{e}$ be a vector such that $\mathbf{e}$ and $\mathbf{h}$ have the same length. Define $\mathbf{e} = [0\ 0\ \cdots\ 0\ 1]^T$. Consider the following linear program:

$$\begin{array}{ll} \text{minimize} & \mathbf{b}^T\mathbf{h} \\ \text{subject to} & \mathbf{D}\mathbf{h} \geq \mathbf{0} \\ & \mathbf{E}\mathbf{h} = \mathbf{0} \\ & \mathbf{e}^T\mathbf{h} = 1 \\ \text{variables:} & \mathbf{h}, \end{array} \qquad \text{(P-Disproof)}$$

and its dual problem (also a linear program):

$$\begin{array}{ll} \text{maximize} & -\gamma \\ \text{subject to} & \mathbf{D}^T\mathbf{y} = \mathbf{b} + \mathbf{E}^T\boldsymbol{\mu} + \mathbf{e}\gamma \\ & \mathbf{y} \geq \mathbf{0} \\ \text{variables:} & \mathbf{y},\ \boldsymbol{\mu}, \gamma. \end{array} \qquad \text{(D-Disproof)}$$

Suppose the Simplex method is used to solve the linear program in (P-Proof) for an information inequality $\mathbf{b}^T\mathbf{h}$ which is not always true. Then the result of (P-Proof) is $-\infty$ from [24, P.344]. However in the linear program in (P-Disproof), we have further fixed $H(X_1, X_2, \ldots, X_n) = 1$ by the extra constraint $\mathbf{e}^T\mathbf{h} = 1$. Together with the inequality constraints in $\mathbf{D}\mathbf{h} \geq 0$, all the elements in any feasible $\mathbf{h}$ (i.e., all the (joint) entropies) are upper bounded by 1. Therefore, $\mathbf{b}^T\mathbf{h}$ is a bounded negative value in the linear program in (P-Disproof).

We have discussed the difficulty of using the optimal $\mathbf{h}^*$, which is obtained by solving the linear program in (P-Disproof), to construct a counterexample. The following theorem states that the optimal $\mathbf{y}^*$ in the dual problem in (D-Disproof) provides us a list of functional dependencies and (conditional) independencies between $\{X_1, X_2, \ldots, X_n\}$. This list gives hints on explicitly constructing a counterexample for disproving $\mathbf{b}^T\mathbf{h} \geq 0$.

*Theorem 3:* Let $[\mathbf{y}^{*T}\ \boldsymbol{\mu}^{*T}]^T$ be the optimal dual solutions for the problem in (D-Disproof). If there exists a joint distribution $P_{X_1, X_2, \ldots, X_n}$ such that its entropy vector $\tilde{\mathbf{h}}$ satisfies

$$\mathbf{y}^{*T}\mathbf{D}\tilde{\mathbf{h}} = \boldsymbol{\mu}^{*T}\mathbf{E}\tilde{\mathbf{h}} = 0, \text{ and } \mathbf{e}^T\tilde{\mathbf{h}} = 1, \tag{27}$$

then $\mathbf{b}^T\tilde{\mathbf{h}} < 0$ and $P_{X_1, \ldots, X_n}$ is a counterexample to disprove $\mathbf{b}^T\mathbf{h} \geq 0$.

*Proof:* Using the KKT conditions, the optimal dual solutions $[\mathbf{y}^{*T}\ \boldsymbol{\mu}^{*T}\ \gamma^*]^T$ for the problem in (D-Disproof) satisfy

$$\mathbf{D}^T\mathbf{y}^* = \mathbf{b} + \mathbf{E}^T\boldsymbol{\mu}^* + \mathbf{e}\gamma^*. \text{ (Stationarity of Lagrangian)}$$

Together with (27), we have

$$\mathbf{b}^T\tilde{\mathbf{h}} = \mathbf{b}^T\tilde{\mathbf{h}} + \boldsymbol{\mu}^{*T}\mathbf{E}\tilde{\mathbf{h}} + \gamma^*\mathbf{e}^T\tilde{\mathbf{h}} - \gamma^* \tag{28}$$

$$= \mathbf{y}^{*T}\mathbf{D}\tilde{\mathbf{h}} - \gamma^* \tag{29}$$

$$= -\gamma^*. \tag{30}$$

From the strong duality of linear programming, i.e., the duality gap is zero in linear programs [25, Theorem 4.4], the optimal value of (P-Disproof) and (D-Disproof) are equal. As the optimal value of (P-Disproof) is negative, $-\gamma^*$ and $\mathbf{b}^T\tilde{\mathbf{h}}$ are both negative so that the theorem is proved. □

In order to satisfy $\mathbf{y}^{*T}\mathbf{D}\tilde{\mathbf{h}} = 0$ in (27), we need to find out the positive elements in $\mathbf{y}^*$. Their corresponding rows in $\mathbf{D}$ tell us the extra equality constraints which are used together with $\mathbf{E}\tilde{\mathbf{h}} = \mathbf{0}$ and $\mathbf{e}^T\tilde{\mathbf{h}} = 1$ to construct a counterexample. Specifically, by solving (P-Disproof) and (D-Disproof), we can obtain a set of equality constraints that give a *sufficient* condition. If we could find a joint distribution of random variables that satisfies the condition, the distribution can be used as a counterexample to disprove the inequality. As we will see, constructing such a joint distribution is not always possible. Let us illustrate the aforementioned disproving process with a few illustrative examples using our AITIP software service (c.f. Section VI).

*Example 4:* A sample output from AITIP:
$AITIP('I(X;Y) <= 0.9H(Y)')$
Not provable by AITIP.
It can be disproved by a probability distribution satisfying all the following Shannon's information measures equal to zero:

$$H(X|Y), H(Y|X). \tag{31}$$

From the above output from AITIP, we can deduce that the counterexample should be chosen as $X = Y$ to disprove $I(X;Y) <= 0.9H(Y)$.

*Example 5:* AITIP can be used to disprove (1):
$AITIP('I(A;B|C,D) + I(B;D|AC) <= I(A;B|D) + I(B;D|A) + H(A) + I(B;D|C)')$
Not provable by AITIP.
It can be disproved by a probability distribution satisfying all the following Shannon's information measures equal to zero:

$$\begin{aligned} &H(A|B,C,D), H(B|A,C,D), H(C|A,B,D), H(D|A,B,C), \\ &I(A;B|C), I(A;B|D), I(A;C|D), I(A;D), I(B;C|A), \\ &I(B;D|A), I(B;D|C), I(C;D|A). \end{aligned} \tag{32}$$

From the above output from AITIP, we can deduce the following counterexample. Let $X, Y$ and $Z$ be three independent binary random variables with entropy equal to 1. Let $(A, B, C, D) = (X \oplus Y, X, Y \oplus Z, Z)$ where $\oplus$ denotes "exclusive or". Then it is easy to check that $I(A;B|D) + I(B;D|A) + H(A) + I(B;D|C) - I(A;B|CD) - I(B;D|CA) = -1 < 0$.

In the above examples, AITIP gives some equality constraints that help us to construct the counterexample for an invalid information inequality. The equality constraints give a sufficient but not necessary condition for an entropy vector to be a counterexample. There are some tricks to construct the example from the output of AITIP. We can consider a set of auxiliary random variables which are mutually independent. By considering the condition entropy in (32), we can obtain some hints about the number of auxiliary random variables. Those mutual information and conditional mutual information in (32) can tell us where "exclusive or" should be used. Of course, the joint entropy of the random variables cannot be zero. Otherwise, the equality constraints are satisfied but the information inequality cannot be disproved.

It is natural to ask: Is it always possible to construct an example from the given equality constraints? Unfortunately, the answer is ‘No’ due to the existence of the non-Shannon-type inequalities. See the following example.

*Example 6:* A sample output from AITIP:
$AITIP('2I(C;D) <= I(A;B)+I(A;C,D)+3I(C;D|A)+I(C;D|B)')$
Not provable by AITIP.
It can be disproved by a probability distribution satisfying all the following Shannon's information measures equal to zero:

$$\begin{aligned}&H(C|A,B,D), I(A;B), I(A;B|C), I(A;C|D), I(A;D|C),\\&I(B;C|D), I(C;D|A), I(C;D|B), I(C;D|A,B). \qquad (33)\end{aligned}$$

In the above example, one should not be able to find a joint distribution $P_{ABCD}$ satisfying (33) with $H(ABCD) \neq 0$. Indeed, it is impossible to find such $P_{ABCD}$ because $2I(C;D) \leq I(A;B)+I(A;C,D)+3I(C;D|A)+I(C;D|B)$ is a non-Shannon-type inequality which is always true for all $P_{ABCD}$ [18]. Therefore, the counterexample does not exist.

It should be noted that there are some heuristics to automate the above process of constructing the probability distributions. In [31], an algorithm is presented to determine whether a given entropy vector $\mathbf{h}$ is *binary* entropic (i.e., whether it corresponds to a probability distribution of some bits), and return the corresponding distribution if it is indeed binary entropic. An extended version of the algorithm solving the generalized problem can be found in [32]. However, the algorithm in [31] requires prior knowledge of the number of auxiliary random variables to use, and both algorithms may fail to return a distribution even with an entropic input, so human insights and interference are still required when using these algorithms. Therefore, these algorithms are not included in our implementation of AITIP, but interested readers can still use them as heuristics to post-process the outputs from our algorithm.

*Remarks:* 1) If we assume that all the quantities in (31) are equal to 0, then we can prove $I(X;Y) \geq 0.9\ H(Y)$, i.e., the opposite of what we wanted to disprove. This property also holds for the inequalities in Examples 5 and 6. This can be seen from Theorem 3. 2) The result from Example 6 can lead us to the following constrained non-Shannon-type inequality.

*Proposition 4:* If all Shannon's information measures in (33) are equal to zero, then $H(A,B,C,D) \leq 0$.

## V. The Smallest Counterexamples and Their Sufficient Conditions

From the previous section, we have seen that the hints for constructing a counterexample to an information inequality can be obtained by first solving a linear program and then inspecting the elemental inequalities corresponding to the positive optimal dual solution. We have also seen that in some special cases, the counterexamples can be directly and systematically obtained through some heuristics [31], [32]. We now explore the *smallest counterexamples* of an (untrue) information inequality as those involving the least number of elemental inequalities. Smaller counterexamples are easier to interpret when trying to construct the joint distribution required to violate the the input inequality, and they tend to reveal more explicitly structure of the distribution. Similarly to the shortest proofs, there often exist multiple smallest counterexamples for an information inequality.

Let us consider the following combinatorial optimization problem:

$$\begin{array}{rl}\text{minimize} & \|[\mathbf{y}^T\ \boldsymbol{\mu}^T]^T\|_0\\ \text{subject to} & \mathbf{D}^T\mathbf{y} = \mathbf{b} + \mathbf{E}^T\boldsymbol{\mu} + \mathbf{e}\gamma^*,\ \mathbf{y} \geq \mathbf{0}\\ \text{variables:} & \mathbf{y},\ \boldsymbol{\mu},\end{array} \qquad (34)$$

where $\|\mathbf{x}\|_0$ is the cardinality or the number of nonzero components in the vector $\mathbf{x}$, and $\gamma^*$ is the optimal solution (also the negative of the optimal value) of (D-Disproof).

Next, we consider the following convex approximation problem to tackle (34):

$$\begin{array}{rl}\text{minimize} & \mathbf{1}^T\mathbf{y} + \|\boldsymbol{\mu}\|_1\\ \text{subject to} & \mathbf{D}^T\mathbf{y} = \mathbf{b} + \mathbf{E}^T\boldsymbol{\mu} + \mathbf{e}\gamma^*,\ \mathbf{y} \geq \mathbf{0}\\ \text{variables:} & \mathbf{y},\ \boldsymbol{\mu}.\end{array} \qquad (35)$$

Furthermore, (35) can be further transformed to the following equivalent problem that is a linear program:

$$\begin{array}{rl}\text{minimize} & \mathbf{1}^T\mathbf{y} + \mathbf{1}^T\mathbf{z}\\ \text{subject to} & \mathbf{D}^T\mathbf{y} = \mathbf{b} + \mathbf{E}^T\boldsymbol{\mu} + \mathbf{e}\gamma^*,\ \mathbf{y} \geq \mathbf{0}\\ & -\mathbf{z} \leq \boldsymbol{\mu} \leq \mathbf{z}\\ \text{variables:} & \mathbf{y},\ \boldsymbol{\mu},\ \mathbf{z}.\end{array} \qquad (36)$$

Given an information inequality $\mathbf{b}^T\mathbf{h} \geq 0$ which is not always true under $\mathbf{D}$ and $\mathbf{E}$, we may want to find a set of conditions $\mathbf{F}$, such that $\mathbf{b}^T\mathbf{h} \geq 0$ is true if both $\mathbf{E}\mathbf{h} = \mathbf{0}$ and $\mathbf{F}\mathbf{h} = \mathbf{0}$ are assumed. In the following, we will show how the set $\mathbf{F}$ can be found by linear programming as a subset of the rows in $\mathbf{D}$. The basic ideas follow from how we disprove an information inequality. The idea is that we are going to find the sufficient conditions which can disprove $\mathbf{b}^T\mathbf{h} < 0$. For the sake of completeness, the details are given as follows.

Let $\mathbf{e}$ be a column vector such that $\mathbf{e}$ and $\mathbf{h}$ have the same length. Define $\mathbf{e} = [0\ 0\ \cdots\ 0\ 1]^T$ due to the last element in $\mathbf{h}$ being the joint entropy $H(X_1, X_2, \ldots, X_n)$. Consider the following linear program (note the difference in the objective function compared with (P-Disproof)):

$$\begin{array}{rl}\text{minimize} & -\mathbf{b}^T\mathbf{h}\\ \text{subject to} & \mathbf{D}\mathbf{h} \geq \mathbf{0}\\ & \mathbf{E}\mathbf{h} = \mathbf{0}\\ & \mathbf{e}^T\mathbf{h} = 1\\ \text{variables:} & \mathbf{h},\end{array} \qquad (37)$$

and its dual problem (also a linear program):

$$\begin{array}{rl}\text{maximize} & -\gamma\\ \text{subject to} & \mathbf{D}^T\mathbf{y} = -\mathbf{b} + \mathbf{E}^T\boldsymbol{\mu} + \mathbf{e}\gamma\\ & \mathbf{y} \geq \mathbf{0}\\ \text{variables:} & \mathbf{y},\ \boldsymbol{\mu}, \gamma.\end{array} \qquad (38)$$

*Theorem 5:* Let $[\mathbf{y}^{*T}\ \boldsymbol{\mu}^{*T}\gamma^*]^T$ be the optimal dual solutions for the problem in (38). Let $\mathbf{y}_i^*$ be the $i$-th entry in $\mathbf{y}^*$ and let $\mathbf{D}_i$ be the $i$-th row in $\mathbf{D}$. For any entropy vector $\tilde{\mathbf{h}}$, if $\mathbf{E}\tilde{\mathbf{h}} = \mathbf{0}$ and

$$\mathbf{D}_i\tilde{\mathbf{h}} = 0 \quad \forall i \in \{j : \mathbf{y}_j^* > 0\}, \qquad (39)$$

then $\mathbf{b}^T\tilde{\mathbf{h}} \geq 0$.

*Proof:* Using the KKT conditions, the optimal dual solutions $[\mathbf{y}^{*T} \ \boldsymbol{\mu}^{*T} \ \gamma^*]^T$ for the problem in (38) satisfy

$$\mathbf{D}^T\mathbf{y}^* = -\mathbf{b} + \mathbf{E}^T\boldsymbol{\mu}^* + \mathbf{e}\gamma^*. \text{ (Stationarity of Lagrangian)}$$

Together with (39), we have

$$\begin{aligned}\mathbf{b}^T\tilde{\mathbf{h}} &= -\mathbf{y}^{*T}\mathbf{D}\tilde{\mathbf{h}} + \boldsymbol{\mu}^{*T}\mathbf{E}\tilde{\mathbf{h}} + \gamma^*\mathbf{e}^T\tilde{\mathbf{h}} \\ &= \gamma^*\mathbf{e}^T\tilde{\mathbf{h}}. \end{aligned} \tag{40}$$

From the strong duality of linear programming, i.e., the duality gap is zero for linear programs, [25, Theorem 4.4], the optimal value of (37) and (38) are equal. As the optimal value of (37) is negative, $\gamma^*$ is positive. Together with $\mathbf{e}^T\tilde{\mathbf{h}} = H(X_1, X_2, \ldots, X_n) \geq 0$, $\mathbf{b}^T\tilde{\mathbf{h}} \geq 0$ from (40), and so the theorem is proved. □

In order to satisfy (39), we need to find out the positive elements in $\mathbf{y}^*$. Their corresponding rows in $\mathbf{D}$ tell us the equality constraints which form the sufficient conditions for the inequality $\mathbf{b}^T\tilde{\mathbf{h}} \geq 0$ to be true.

## VI. Parallel Computation With ADMM

It is well known that the simplex method always looks for vertex solutions (solutions at the corners of the feasibility region). In our problem, at the vertices most constraints are redundant, and therefore the dual solutions (that we use to construct proofs/disproofs) obtained from the simplex method are often highly sparse, making the proofs and disproofs concise and elegant, while numerical algorithms like interior-point method do not guarantee the sparsity of the dual solutions. This is the primary reason why the simplex method is used in the existing software implementations like [7] and [8].

It was shown by Spielman and Tang [33] that the simplex algorithm usually has polynomial running time, but in practice it is sometimes less efficient compared to other linear program algorithms like interior-point method. In our particular problem, when we are solving gigantic, highly sparse and highly degenerate linear programs, the performance of the simplex algorithm deteriorates. In our experience, solving the linear programs in the framework for a relatively large number of random variables can be extremely slow or even intractable, and therefore a more efficient algorithm is needed. In this section, we develop an efficient and parallelizable algorithm based on ADMM for the linear programming framework. As we shall see, the ADMM-based algorithmic framework improves the process of searching for the sparse dual solution by iterative updates that have closed-form solutions, thus enabling highly parallelizable and distributed computation.

It should be noted that, like interior-point method, ADMM is a numerical optimization algorithm, and the optimal dual solution is often dense. Therefore, in our implementation we are using a simplex-based "crossover" (also known as "purification" or "the BASIC procedure" in some literature) technique to solve for vertex solutions with sparse dual vectors. This is a well-established and widely used technique in numerical optimization (see [34]–[36]), and it is built-in in many commercial linear program solvers.

As with other numerical algorithms, small floating-point rounding errors and numerical precision issues are unavoidable, and in practice, any software implementation should be tuned to the desired numerical accuracy. In our numerical experiments as well as our online service (more details in the following sections), the standard double precision arithmetic has been sufficient.

Before we proceed, we would like to address that our framework is designed as a general prover to prove or disprove information inequalities, but for some specific inequalities, by exploiting the problem structures, it is possible to design problem-specific algorithms with superior performance. One example is [11] where the problem can be viewed as proving an information inequality with 16 random variables. Such problem size is intractable for existing software implementations. Fortunately, by exploiting the symmetric structure in the specific problem, many constraints and variables can be shown to be redundant and thus can be removed from the linear program, leaving only 76 variables and $6,152$ constraints. Similarly, in [37], [38], the minimal set of elemental inequalities under functional dependencies and full conditional independence structures is characterized. With some specific types of user-defined constraints (the $\mathbf{E}$ matrix), the number of elemental inequalities to use (the size of the $\mathbf{D}$ matrix) can be significantly reduced, making the resulting linear programs easier to solve.

### A. Generic ADMM Algorithm

Consider the optimization problem

$$\begin{array}{ll} \text{minimize} & f(\mathbf{x}) + g(\mathbf{y}) \\ \text{subject to} & \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{y} = \mathbf{c} \\ \text{variables:} & \mathbf{x}, \mathbf{y}, \end{array} \tag{41}$$

with $\rho$-augmented Lagrangian

$$L_\rho = f(\mathbf{x}) + g(\mathbf{y}) + \boldsymbol{\lambda}^T(\mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{y} - \mathbf{c}) + \frac{\rho}{2}||\mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{y} - \mathbf{c}||^2,$$

where $\rho$ is a hyperparameter that we can choose.[5] A generic ADMM algorithm is given in Algorithm 1.

**ALGORITHM 1** Generic ADMM Algorithm

```
repeat
    1. x-update: $\mathbf{x}^{k+1} = \arg\min_x\{L_\rho(\mathbf{x}, \mathbf{y}^k, \boldsymbol{\lambda}^k)\}$
    2. y-update: $\mathbf{y}^{k+1} = \arg\min_y\{L_\rho(\mathbf{x}^{k+1}, \mathbf{y}, \boldsymbol{\lambda}^k)\}$
    3. λ-update: $\boldsymbol{\lambda}^{k+1} = \boldsymbol{\lambda}^k + \rho(\mathbf{A}\mathbf{x}^{k+1} + \mathbf{B}\mathbf{y}^{k+1} - \mathbf{c})$
until Convergence;
```

The ADMM algorithm and its convergence analysis have been introduced with great details in [39] and [40, 4.4]. Here we summarize the convergence result in Theorem 6. The proof can be found in [39, Appendix A].

*Assumption 1:* The objective functions $f$ and $g$ are closed proper convex.

*Assumption 2:* The unaugmented Lagrangian $L_0$ has a saddle point.

*Theorem 6:* Under Assumption 1 and 2, as $k \to \infty$, Algorithm 1 satisfies the following:

[5] In all our numerical experiments in Section VII and the online web service AITIP (cf. Section VIII), we use $\rho = 2$.

1) *Residual Convergence*: $\mathbf{A}\mathbf{x}^k + \mathbf{B}\mathbf{y}^k - \mathbf{c} \to \mathbf{0}$.
2) *Objective Convergence*: $f(\mathbf{x}^k) + g(\mathbf{y}^k) \to p^*$, where $p^*$ is the optimal objective value.
3) *Dual Variable Convergence*: $\boldsymbol{\lambda}^k \to \boldsymbol{\lambda}^*$, where $\boldsymbol{\lambda}^*$ is the optimal dual solution.

### *B. Problem Reformulation*

Consider the following linear program

$$\begin{array}{ll} \text{minimize} & \mathbf{b}^T\mathbf{h} \\ \text{subject to} & \mathbf{0} \le \mathbf{D}\mathbf{h} \le \mathbf{1} \\ & \mathbf{E}\mathbf{h} = \mathbf{0} \\ \text{variables:} & \mathbf{h}, \end{array} \qquad \text{(P-Merge)}$$

and its dual problem

$$\begin{array}{ll} \text{maximize} & -\mathbf{1}^T\boldsymbol{\lambda}_2 \\ \text{subject to} & \mathbf{b} - \mathbf{D}^T\boldsymbol{\lambda}_1 + \mathbf{D}^T\boldsymbol{\lambda}_2 + \mathbf{E}^T\boldsymbol{\mu} = \mathbf{0} \\ & \boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2 \ge \mathbf{0} \\ \text{variables:} & \boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2, \boldsymbol{\mu}, \end{array} \qquad \text{(D-Merge)}$$

we have the following results.

*Lemma 7:* If $\mathbf{0} \le \mathbf{D}\mathbf{h} \le \mathbf{1}$, then $\mathbf{h}$ is box-bounded (i.e., all elements of $\mathbf{h}$ are bounded above and below).

*Proof:* Let $\{X_1, X_2, \cdots, X_n\}$ be the random variables involved in the information inequality to prove and let $\mathcal{N} = \{1, 2, \cdots, n\}$. For all $i$ in $\mathcal{N}$, $H(X_i)$ can be written as the sum of $n$ elemental inequalities in the form of (6) and (7).

For example, when $n = 4$, $H(X_1)$ can be written as

$$\begin{aligned} H(X_1) &= H(X_1|X_2X_3X_4) + I(X_1;X_2X_3X_4) && (42) \\ &= H(X_1|X_2X_3X_4) + I(X_1;X_2|X_3X_4) \\ &\quad + I(X_1;X_3X_4) && (43) \\ &= H(X_1|X_2X_3X_4) + I(X_1;X_2|X_3X_4) \\ &\quad + I(X_1;X_3|X_4) + I(X_1;X_4), && (44) \end{aligned}$$

where in (42) we use the Proposition 2.19 in [24], and in (43) and (44) we use the chain rule for mutual information successively [24, Proposition 2.26].

The right hand side of (44) is the sum of 4 rows in $\mathbf{D}\mathbf{h}$. Since $\mathbf{0} \le \mathbf{D}\mathbf{h} \le \mathbf{1}$, we have $0 \le H(X_1) \le 4$. The same argument can be applied to all random variables, and therefore the optimization variable $\mathbf{h}$ is box-bounded. □

*Theorem 8:* Both (P-Merge) and (D-Merge) have finite optimal solutions (i.e., they are "solvable").

*Theorem 9:* A set of optimal solutions to (D-Merge), $\left[\boldsymbol{\lambda}_1^{*T}\ \boldsymbol{\lambda}_2^{*T}\ \boldsymbol{\mu}^{*T}\right]^T$, can be used to construct a proof or counterexample to the information inequality $\mathbf{b}^T\mathbf{h} \ge 0$.

These two theorems are essential in our AITIP algorithm, as they show that any information inequality (assuming it is not non-Shannon-type) can be proved or disproved by solving *a single* linear program that is guaranteed to be "solvable". The implication is that a unified computational approach to prove or disprove an information inequality can be developed. The proofs are given below.

*Proof of Theorem 9:* Since strong duality holds in linear programming, (D-Merge) has finite optimum if and only if (P-Merge) also has finite optimum [25, Table 4.2], so we only need to prove the feasibility and below-boundedness of (P-Merge).

Now, (P-Merge) is obviously feasible, as the trivial solution $\mathbf{h}^* = \mathbf{0}$ is a feasible solution. The primal constraint $\mathbf{0} \le \mathbf{D}\mathbf{h} \le \mathbf{1}$ implies that the optimization variable $\mathbf{h}$ is box-bounded (Lemma 7), therefore the objective value $p^*$ is bounded from below. □

*Proof of Theorem 8:* When the information inequality is true, it is easy to see that in both (P-Proof) and (P-Merge), we have $\mathbf{h}^* = \mathbf{0}$ and $p^* = 0$. Thus the extra constraint in (P-Merge), $\mathbf{D}\mathbf{h} \le \mathbf{1}$, is inactive, i.e., redundant, and therefore $\boldsymbol{\lambda}_2^* = \mathbf{0}$. Using this together with the stationarity of Lagrangian

$$\mathbf{b} - \mathbf{D}^T\boldsymbol{\lambda}_1 + \mathbf{D}^T\boldsymbol{\lambda}_2 + \mathbf{E}^T\boldsymbol{\mu} = \mathbf{0}, \qquad (45)$$

for any primal feasible $\mathbf{h}$, we have

$$\begin{aligned} \mathbf{b}^T\mathbf{h} &= \boldsymbol{\lambda}_1^*\mathbf{D}\mathbf{h} - \boldsymbol{\lambda}_2^*\mathbf{D}\mathbf{h} + \boldsymbol{\nu}^*\mathbf{E}\mathbf{h} \\ &= \boldsymbol{\lambda}_1^*\mathbf{D}\mathbf{h} \\ &\ge 0, \end{aligned}$$

which can be viewed as the certificate of the nonnegativity of $\mathbf{b}^T\mathbf{h}$ for any feasible $\mathbf{h}$, and therefore a proof can be constructed similarly to what has been illustrated in Example 1.

If the information inequality is not provable, similarly to the disproof construction in Example 4 and Example 5, the solutions $\{\boldsymbol{\lambda}_1^*, \boldsymbol{\lambda}_2^*, \boldsymbol{\mu}^*\}$ can be used to construct a counterexample. Specifically, assuming that the input inequality is not provable, if there exists a valid entropy vector $\tilde{\mathbf{h}}$ satisfying

$$\boldsymbol{\lambda}_1^{*T}\mathbf{D}\tilde{\mathbf{h}} = \boldsymbol{\lambda}_2^{*T}(\mathbf{D}\tilde{\mathbf{h}} - \mathbf{1}) = \boldsymbol{\mu}^{*T}\mathbf{E}\tilde{\mathbf{h}} = 0, \qquad (46)$$

then $\tilde{\mathbf{h}}$ can be used to construct a counterexample to disprove the inequality using (46) together with (45), as

$$\begin{aligned} \mathbf{b}^T\tilde{\mathbf{h}} &= \mathbf{b}^T\tilde{\mathbf{h}} - \boldsymbol{\lambda}_1^{*T}\mathbf{D}\tilde{\mathbf{h}} + \boldsymbol{\mu}^{*T}\mathbf{E}\tilde{\mathbf{h}} \\ &= -\mathbf{D}^T\boldsymbol{\lambda}_2^* \\ &= -\mathbf{1}^T\boldsymbol{\lambda}_2^*. \end{aligned}$$

Since we have assumed the inequality is false, we know that the optimal objective value of (D-Merge), $-\mathbf{1}^T\boldsymbol{\lambda}_2^*$, should be strictly negative, thus $\mathbf{b}^T\tilde{\mathbf{h}} = -\mathbf{1}^T\boldsymbol{\lambda}_2^* < 0$. Therefore, such an $\tilde{\mathbf{h}}$ can be used to construct a counterexample to disprove the inequality. □

Now consider the following linear program

$$\begin{array}{ll} \text{minimize} & \mathbf{b}^T\mathbf{h} \\ \text{subject to} & \mathbf{D}\mathbf{h} - \mathbf{u} = \mathbf{0} \\ & \mathbf{D}\mathbf{h} + \mathbf{v} = \mathbf{1} \\ & \mathbf{E}\mathbf{h} = \mathbf{0} \\ & \mathbf{u}, \mathbf{v} \ge \mathbf{0} \\ \text{variables:} & \mathbf{h}, \mathbf{u}, \mathbf{v}, \end{array} \qquad \text{(P-ADMM)}$$

and its dual problem

$$\begin{array}{ll} \text{maximize} & -\mathbf{1}^T\boldsymbol{\nu}_2 \\ \text{subject to} & \mathbf{b} + \mathbf{D}^T\boldsymbol{\nu}_1 + \mathbf{D}^T\boldsymbol{\nu}_2 + \mathbf{E}^T\boldsymbol{\mu} = \mathbf{0} \\ & \boldsymbol{\nu}_1 \le \mathbf{0}, \boldsymbol{\nu}_2 \ge \mathbf{0}, \\ \text{variables:} & \boldsymbol{\nu}_1, \boldsymbol{\nu}_2, \boldsymbol{\mu}, \end{array} \qquad \text{(D-ADMM)}$$

where (P-ADMM) is obtained by introducing slack variables $\mathbf{u}$ and $\mathbf{v}$ to (P-Merge). Comparing (D-ADMM) and (D-Merge), it is easy to see that if we have a set of optimal solutions to (D-ADMM), $\left[\boldsymbol{\nu}_1^{*T}\ \boldsymbol{\nu}_2^{*T}\ \boldsymbol{\mu}^{*T}\right]^T$, we also obtain the optimal solutions to (D-Merge) as $\left[-\boldsymbol{\nu}_1^{*T}\ \boldsymbol{\nu}_2^{*T}\ \boldsymbol{\mu}^{*T}\right]^T$ "for free". Therefore, solving (P-ADMM) and (D-ADMM) is enough to prove or disprove any given Shannon-type information inequality.

Finally, we rewrite (P-ADMM) to the following form so that it is more compact and easier to work with

$$\begin{aligned} &\text{minimize} && \mathbf{b}^T\mathbf{h} \\ &\text{subject to} && \mathbf{B}\mathbf{h}+\mathbf{y}=\mathbf{c} \qquad \text{(P-ADMM)} \\ &\text{variables:} && \mathbf{h},\mathbf{y}, \end{aligned}$$

where $\mathbf{B}=\left[\mathbf{D}^T\ \mathbf{D}^T\ \mathbf{E}^T\right]^T$, $\mathbf{y}=\left[-\mathbf{u}^T\ \mathbf{v}^T\ \mathbf{0}^T\right]^T$, $\mathbf{u},\mathbf{v}\geq 0$ and $\mathbf{c}=\left[\mathbf{0}^T\ \mathbf{1}^T\ \mathbf{0}^T\right]^T$.

*Remark:* In Section III and V respectively, in order to obtain the "shortest proofs" and the "smallest counterexamples", we construct two linear programs, (17) and (36), to approximately solve the $l_0$ norm optimization problems. The dual problems of those two linear programs can also be reformulated into the form of (P-ADMM), and therefore our main algorithm to be introduced in the next subsection is applicable in these two cases.

### C. AITIP Algorithm

The $\rho$-augmented Lagrangian of (P-ADMM) is

$$L_\rho=\mathbf{b}^T\mathbf{h}+\boldsymbol{\nu}^T(\mathbf{B}\mathbf{h}+\mathbf{y}-\mathbf{c})+\frac{\rho}{2}||\mathbf{B}\mathbf{h}+\mathbf{y}-\mathbf{c}||^2,$$

where $\boldsymbol{\nu}$ is the Lagrangian multiplier of the constraint $\mathbf{B}\mathbf{h}+\mathbf{y}=\mathbf{c}$. With ADMM, we can solve (P-ADMM) using Algorithm 2, which we call the *AITIP Algorithm.*

**ALGORITHM 2** AITIP Algorithm

```
repeat
    1. h-update: $\mathbf{h}^{k+1}=\arg\min_h\{L_\rho(\mathbf{h},\mathbf{y}^k,\boldsymbol{\nu}^k)\}$
    2. y-update: $\mathbf{y}^{k+1}=\arg\min_y\{L_\rho(\mathbf{h}^{k+1},\mathbf{y},\boldsymbol{\nu}^k)\}$
    3. ν-update: $\boldsymbol{\nu}^{k+1}=\boldsymbol{\nu}^k+\rho(\mathbf{B}\mathbf{h}^{k+1}+\mathbf{y}^{k+1}-\mathbf{c})$
until Convergence;
```

Recall that, as stated in Theorem 6, an ADMM-based algorithm converges under the following two mild assumptions:

1) the objective functions $f$ and $g$ as in Algorithm 1 are closed proper convex.
2) the unaugmented Lagrangian $L_0$ has a saddle point.

Assumption 1) is obviously true since the objective function is linear in our case. Because our problem is a linear program, assumption 2) is equivalent to saying that both the primal and dual problems are solvable, which has been proved in Theorem 8. Therefore, the AITIP Algorithm does converge.

The $\mathbf{h}$-update in Step 1 is an unconstrained quadratic program, so we can get the following closed-form solution:

$$\mathbf{h}^{k+1}=-\frac{1}{\rho}(\mathbf{B}^T\mathbf{B})^{-1}(\mathbf{b}+\mathbf{B}^T\boldsymbol{\nu}^k+\rho\mathbf{B}^T\mathbf{y}^k-\rho\mathbf{B}^T\mathbf{c}). \qquad (\mathbf{h}\text{-update})$$

*Remark:* In ($\mathbf{h}$-update) we use the inversion of $\mathbf{B}^T\mathbf{B}$ only for showing the closed-form solution. In practice the inversion of a large matrix is both inefficient and numerically unstable. Instead, we apply the Cholesky factorization on $\mathbf{B}^T\mathbf{B}$ to get a lower-triangular matrix $\mathbf{L}$ such that $\mathbf{B}^T\mathbf{B}=\mathbf{L}\mathbf{L}^T$, and directly solve the linear system via back substitution.

Since $\mathbf{y}=\left[-\mathbf{u}^T\ \mathbf{v}^T\ \mathbf{0}^T\right]^T$, the $\mathbf{y}$-update in Step 2 can be decomposed into $\mathbf{u}$-update and $\mathbf{v}$-update. The $\mathbf{u}$-update is a constrained quadratic program, but luckily the KKT system can be directly solved, giving us the closed-form solution

$$\mathbf{u}^{k+1}=(\mathbf{D}\mathbf{h}^{k+1}+\frac{1}{\rho}\boldsymbol{\nu}_u^k)_+, \qquad (\mathbf{u}\text{-update})$$

and similarly for $\mathbf{v}$-update,

$$\mathbf{v}^{k+1}=(\mathbf{1}-\mathbf{D}\mathbf{h}^{k+1}-\frac{1}{\rho}\boldsymbol{\nu}_v^k)_+. \qquad (\mathbf{v}\text{-update})$$

In ($\mathbf{u}$-update) and ($\mathbf{v}$-update) above, $\boldsymbol{\nu}_u^k$ is the upper half of $\boldsymbol{\nu}^k$, $\boldsymbol{\nu}_v^k$ is the lower half of $\boldsymbol{\nu}^k$, and $(\mathbf{x})_+=\max\{\mathbf{x},\mathbf{0}\}$ where $\max$ gives the elementwise maximum.

We can see that every sub-problem in each step in the AITIP Algorithm has a closed-form solution, thus it can be solved very efficiently. As we mentioned at the beginning of the section, after the convergence of the AITIP Algorithm, a "crossover" step is required to obtain sparse dual solutions.

### D. Efficient Matrix Factorization in $\mathbf{h}$-Update

In the $\mathbf{h}$-update, we may need to factorize a large matrix $\mathbf{B}^T\mathbf{B}$ at each iteration. Let $n$ be the number of random variables involved in the inequality and $k$ be the dimension of the $h$ vector, it is shown in [6] that $k=2^n-1$. Since $\mathbf{B}^T\mathbf{B}\in\mathbb{R}^{k\times k}$, its dimension grows exponentially with the number of random variables. Therefore, when $n$ is large, the factorization of $\mathbf{B}^T\mathbf{B}$ is the computational bottleneck of the AITIP Algorithm.

Since $\mathbf{B}$ is fixed throughout the algorithm, an obvious approach is to factorize it right at the beginning of the algorithm, and directly use the factorization in all subsequent iterations. However, this is still not ideal as Cholesky factorization generally has a time complexity of $O(k^3)$. In this subsection, we develop an efficient method to compute the factorization.

Recall that $\mathbf{B}=\left[\mathbf{D}^T\ \mathbf{D}^T\ \mathbf{E}^T\right]^T$, where the only problem-dependent component is $\mathbf{E}$. In other words, different input inequalities can share the same $\mathbf{B}$ matrix, if they involve the same number of random variables $n$, and they do not contain problem-specific constraints ($\mathbf{E}$ does not exist). Therefore, we can pre-factorize the $\mathbf{B}^T\mathbf{B}$ matrix for different $n$ values and cache the factorization on disk. If the input inequality has no problem-specific constraint, the factorization can be loaded directly from disk to speed up the computation.

If the input inequality does contain user-constraints, we can still obtain the corresponding Cholesky factorization from the cache. Let $\mathbf{e}_i^T$ be the $i$-th row of the matrix $\mathbf{E}$, and let the number of problem-specific constraints be $q$, i.e., $\mathbf{E}\in\mathbb{R}^{q\times k}$. Let $\tilde{\mathbf{B}}$ be $\left[\mathbf{D}^T\ \mathbf{D}^T\right]^T$, we have

$$\mathbf{B}^T\mathbf{B}=\tilde{\mathbf{B}^T}\tilde{\mathbf{B}}+\sum_{i=1}^{q}\mathbf{e}_i\mathbf{e}_i^T,$$

TABLE I

AVERAGE RUNNING TIME IN SECONDS FOR DIFFERENT ALGORITHMS WHEN PROVING RANDOMLY GENERATED INEQUALITIES INVOLVING 10 TO 14 RANDOM VARIABLES

| n | Gurobi Simplex (1 CPU Core) | Gurobi IP (12 CPU Cores) | AITIP (1 CPU Core) | AITIP (12 CPU Cores) | AITIP (GPU) | Crossover (1 CPU Core) |
|---|---|---|---|---|---|---|
| 10 | 1.88 (10) | 0.73 (10) | 1.99 (10) | 0.55 (10) | 0.30 (10) | 1.22 (10) |
| 11 | 20.08 (10) | 3.61 (10) | 11.94 (10) | 2.20 (10) | 0.91 (10) | 5.44 (10) |
| 12 | 269.32 (10) | 10.51 (10) | 49.13 (9) | 6.62 (10) | 8.06 (10) | 35.09 (10) |
| 13 | 3409.55 (3) | 51.33 (10) | 231.81 (6) | 29.55 (10) | 11.81 (10) | 248.46 (10) |
| 14 | (0) | 296.43 (10) | 522.82 (5) | 178.06 (10) | 54.06 (10) | 1430.12 (10) |

i.e., $\mathbf{B}^T\mathbf{B}$ can be constructed from $\widetilde{\mathbf{B}^T}\tilde{\mathbf{B}}$ via a series of rank-1 updates. Since we have the cache of factorization of $\widetilde{\mathbf{B}^T}\tilde{\mathbf{B}}$, the factorization of $\mathbf{B}^T\mathbf{B}$ can be efficiently obtained with complexity $O(k^2q)$, as discussed in [41]. In practice, the number of problem-specific constraints is usually small, thus $q \ll k$, and therefore this method is much faster than running the factorization from scratch. The decomposition of $\mathbf{B}^T\mathbf{B}$ in this form is in fact instrumental in engineering a scalable software-as-a-service (cf. Section VIII), which is an online service usable by anyone interested to solve their specific problems. The online cache of factorization of $\mathbf{B}^T\mathbf{B}$ can simply be shared by all the users.

## VII. NUMERICAL EXPERIMENT

The proposed AITIP Algorithm is implemented in C++. Gurobi [42], a state-of-the-art commercial linear program solver, is used in the crossover step upon the convergence of our AITIP algorithm. The original simplex-based algorithm is also implemented using Gurobi as a baseline for comparison.

Since all the sub-problems in the AITIP Algorithm have closed-form solutions, the algorithm is essentially a series of (sparse) linear algebra operations, which can be easily parallelized to boost the performance. To fully unleash the power of this algorithm, it is also implemented using the CUDA toolkits (cuSPARSE and cuBLAS) to be executed on the GPU [43]. In the following subsections, all the CPU computations are done on a GNU/Linux machine with two 6-core Intel® Xeon® CPUs, and the GPU computations are done on a machine with one Nvidia® V100 GPU installed.

### *A. Randomly Generated Information Inequalities*

For each $n$ value between 10 and 14, we randomly generate 10 inequalities and attempt to prove them with the Gurobi solver and our proposed AITIP algorithm. In Gurobi, the simplex method uses a single CPU core only, while the interior-point method would use all the CPU cores (12 on our test machine) by default, so we test AITIP in both CPU settings to make the comparison fair. The results are summarized in Table I below.

Each column in the table contains the average running time of an algorithm in seconds. In each table cell, the integer inside the parentheses is the number of generated inequalities (out of 10) that the algorithm has managed to prove or disprove within 3600 seconds time limit, and the number outside of the parentheses is the average solving time for those inequalities proved in 3600 seconds. For example, the top left cell means that for $n = 10$, the Simplex solver in Gurobi can solve all the 10 generated inequalities within the 3600 seconds limit, and the average running time is 1.88 seconds. Note that "Crossover" is merely a post-processing step to produce sparse results (see the last subsection), and it cannot prove or disprove inequalities by itself. All listed algorithms except "Gurobi Simplex" need to undergo the crossover step to generate short proofs and disproofs.

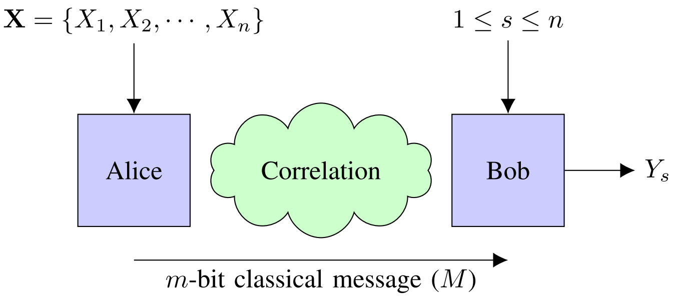

$\mathbf{X} = \{X_1, X_2, \cdots, X_n\}$
$1 \le s \le n$
Alice
Correlation
Bob
$Y_s$
$m$-bit classical message ($M$)

Fig. 4. IC principle illustrated as a cryptographic game.

From Table I, we can see that AITIP with crossover consistently outperforms both linear programming algorithms in Gurobi when running under the same CPU settings. With access to a GPU, AITIP with crossover is significantly faster and more robust than any other algorithms.

### *B. Application in Proving the Information Causality Inequality*

In this subsection, we demonstrate how our proposed framework can be used to prove an important inequality in quantum communication, and show the superior performance of our AITIP algorithm compared to the existing software implementations.

In [44], the authors proposed a physical principle called Information Causality (IC). The principle can be best understood as a game illustrated in Fig. 4. Alice receives an $n$-bit string $\mathbf{X} = \{X_1, X_2, \cdots, X_n\}$, and Bob receives a random integer $s$ between 1 and $n$ from a separate location. Alice is then allowed to send an $m$ bit classical message ($M$) to Bob. Now Bob's task is to guess the value of the $s$-th bit of $\mathbf{X}$ (i.e.,$X_s$) as $Y_s$, with the help of the message from Alice and a pre-established correlation between them (represented by the cloud-shaped icon). The correlation can be of any form (classical or quantum), but it is assumed to be *no-signaling*. Here no-signaling means that the correlation contains no information of $\mathbf{X}$. Formally, let $B$ be what Bob can measure from the correlation, the no-signalling condition ensures that $I(\mathbf{X}; B) = 0$.

The IC principle states that, *information gain that Bob can reach about previously unknown to him data set of Alice, by using all his local resources and $m$ classical bits communicated by Alice, is at most $m$ bits [44].* More precisely, denote Bob's guess as $Y_i$, and define the efficiency of the strategy as

$$\eta \equiv \sum_{i=1}^{n} I(X_i; Y_i | s = i),$$

we have $\eta \leq m$.

Recently in [45], the authors generalized the idea of causal structure graph to quantum systems, and using their new techniques they found a tighter version of the original IC inequality (eq.(10) in [45]), which holds under both classical and quantum cases. The new inequality is

$$\sum_{i=1}^{n} I(X_i; Y_i, M) + \sum_{i=2}^{n} I(X_1; X_i | Y_i, M) \leq$$
$$H(M) + \sum_{i=1}^{n} H(X_i) - H(X_1, \cdots, X_n).$$

In the classical case, the inequality can be proved using the following two constraints [45, Supplementary Note]:

$$I(X_{\mathcal{N}}; \alpha) = 0,$$
$$I(X_{\mathcal{N}}; Y_{\mathcal{N}} | M, \alpha) = 0,$$

where $\mathcal{N} = \{1, 2, \cdots, n\}$ and $\alpha$ is the classical analog of the quantum correlation.

When $n = 6$, the inequality to prove contains $14$ random variables, and the corresponding linear program has $16,383$ optimization variables and $372,752$ constraints, which is far beyond the capability of the existing software implementations of the linear programming framework [7], [8]. For example, XITIP [8] failed to solve the problem in $48$ hours even when running on a powerful Intel® Xeon® CPU. In comparison, the implementation of our proposed AITIP algorithm solves the problem much faster. Using the exact same setup as in subsection VII-A, the identity is proved in $1,855.86$ seconds on a CPU and $68.49$ seconds on a GPU. The computer-generated analytic proof contains $64$ elemental inequalities, which is quite concise compared to the problem size.

## VIII. Further Discussions

In this section we summarize our key contributions and suggest a few possible future directions to advance our linear programming framework for proving and disproving information inequalities.

### A. Sparse Dual Solutions Without Crossover

As shown in Table I, in the AITIP algorithm, most of the running time are used in the post-processing phase (crossover), which is intrinsically slow as it is based on the simplex method. The sole purpose of this post-processing phase is to obtain sparse solutions which would lead to shorter proofs or counterexamples, and therefore it would be of interest if we could directly solve the linear programs and get sufficiently sparse dual solutions without resorting to the simplex method.

In Section III and V, the $l_1$-norm heuristic is used to get the shortest proofs and the smallest counterexamples, but from our numerical observations and experience, the simplex method is still needed even with the $l_1$-norm term in the objective function, as otherwise the dual solutions are not sparse enough. A plausible reason is that $l_1$-norm is a poor approximation to the vector cardinality for those instances, so an alternative would be to use a better (possibly non-convex) function to approximate $\|\mathbf{x}\|_0$. Recently the authors of [46] took a similar approach, where they used a non-convex function to approximate the cardinality function and then solved the non-convex optimization problem using the majorization-minimization technique.

### B. Automatically Identify Structures

In Section VI, we point out that by exploiting certain problem structures, it is possible to reduce the linear program size and therefore improve the performance significantly. We would like to address that such structures do not always exist in general information inequalities, and even when the inequalities do have these required structures, careful inspection by trained human eyes is required to identify and exploit them. Because of this, it is of interest if AITIP can be extended to automatically identify and exploit these hidden structures to reduce the problem size and thus improve its performance.

As shown in Example 5, when the given inequality is not true in general, AITIP would provide hints which can then be used to construct counterexamples to disprove the inequality. However, careful human intervention is still required to fill the gap between provided hints and the actual counterexamples (e.g. constructing the chain of binary random variables as in Example 5). We did not explore this direction in this paper, but it is possible to design an algorithm that does the construction automatically and thus making the framework fully automated. Indeed, this paper can be viewed as a case study of the "automated reasoning by convex optimization" approach [47] to enable automated problem solving in the field of artificial intelligence.

### C. Scalable Computation and Software-as-a-Service

In subsection VI-D we see that part of the computation in the AITIP algorithm (the factorization of the $\mathbf{B}^T\mathbf{B}$ matrix) can be done a priori, caching its values to improve the performance. It is natural to implement this algorithm as a cloud computing service, due to the following reasons:

- The $\mathbf{B}^T\mathbf{B}$ matrices can take quite some time to be decomposed, especially on a small desktop computer, and it is inefficient since each user would have to generate the factorization matrices before using them. With a cloud service, a single set of decomposed matrices can be shared by different users solving their own problems, making the system more scalable to tackling many problem instances.
- For large $n$ values, the cached factorization matrices can take up a few gigabytes of disk space, which makes it impractical to be used in desktop computers. When used in a cloud computing service, the matrices can be easily

stored and shared among many different users due in part to the nature of cloud computing and the massive storage on cloud servers.

Existing software packages, namely ITIP in [7] and XITIP [8], are designed to run on a single desktop computer, which typically cannot handle large problems. These software packages also have software dependency issues, operating system and backward compatibility requirements that limit long-term usage. In view of this, we have developed a Software-as-a-Service platform using cloud computing to automate the tasks of finding proofs or counterexamples using the aforementioned algorithmic framework. The platform, which we call AITIP, is available at `https://aitip.org`. Since the service is currently running on a single-CPU machine on Google Cloud, we have to limit the number of random variables $n$ to be $\leq 12$, so that the available computational resources can be shared by all users in a more reasonable way. For $n \leq 12$, the input inequalities can typically be proved or disproved in a few minutes. To prove inequalities involving more random variables, the users can download the source code available at `https://github.com/convexsoft/AITIP` and run it on their own machines.

## IX. Conclusion

Information theoretic inequalities play a crucial role in various fields, therefore it is important to develop an automated prover to automate the task of proving and disproving them. In this paper we presented a novel theoretical framework for proving and disproving information inequalities using linear programming. By utilizing the $l_1$-norm heuristic for sparse optimization, our proposed framework can also construct the shortest proofs and the smallest counterexamples. The sufficient conditions to manipulate the given inequality to become true can also be identified by the framework. Lastly, we proposed a scalable and efficient algorithm based on the alternating direction method of multipliers to make the proving and disproving of large scale information inequalities possible, and we developed an online web service based on the AITIP algorithm as a proof of concept.

## References

[1] S.-W. Ho, C. W. Tan, and R. W. Yeung, "Proving and disproving information inequalities," in *Proc. IEEE Int. Symp. Inf. Theory*, Jun. 2014, pp. 2814–2818.

[2] C. W. Tan, S.-W. Ho, S. Lint, and R. W. Yeung, "Finding the capacity of next-generation networks by linear programming," in *Proc. IEEE Int. Conf. Commun. Syst.*, Nov. 2014, pp. 192–196.

[3] L. Ling, C. W. Tan, S.-W. Ho, and R. W. Yeung, "Scalable automated proving of information theoretic inequalities with proximal algorithms," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jul. 2019, pp. 1382–1386.

[4] H. Wang, "Computer theorem proving and artificial intelligence," in *Automated Theorem Proving: After 25 Years, American Mathematical Society, Contemporary Mathematics*, vol. 29. Dordrecht, The Netherlands: Springer, 1984, pp. 49–70.

[5] T. C. Hales, "Linear programs for the Kepler conjecture," in *Mathematical Software—ICMS* (Lecture Notes in Computer Science), vol. 6327, K. Fukuda, J. Hoeven, M. Joswig, and N. Takayama, Eds. Berlin, Germany: Springer, 2010.

[6] R. W. Yeung, "A framework for linear information inequalities," *IEEE Trans. Inf. Theory*, vol. 43, no. 6, pp. 1924–1934, Nov. 1997.

[7] R. W. Yeung and Y. O. Yan. (1999). *Information Theoretic Inequality Prover (ITIP), MATLAB Program Software Package*. [Online]. Available: http://home.ie.cuhk.edu.hk/~ITIP

[8] R. Pulikkoonattu and S. Diggavi. (2006). *ITIP-Based C Program Software Package*. [Online]. Available: http://xitip.epfl.ch

[9] L. Csirmaz. (2016). *Minitip—A MINimal Information Theoretic Inequality Prover*. [Online]. Available: https://github.com/lcsirmaz/minitip

[10] C. T. Li. (2020). *Psitip—Python Symbolic Information Theoretic Inequality Prover*. [Online]. Available: https://github.com/cheuktingli/psitip

[11] C. Tian, "Characterizing the rate region of the (4,3,3) exact-repair regenerating codes," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 5, pp. 967–975, May 2014.

[12] R. W. Yeung and Z. Zhang, "Distributed source coding for satellite communications," *IEEE Trans. Inf. Theory*, vol. 45, no. 4, pp. 1111–1120, May 1999.

[13] X. Yan, R. W. Yeung, and Z. Zhang, "An implicit characterization of the achievable rate region for acyclic multisource multisink network coding," *IEEE Trans. Inf. Theory*, vol. 58, no. 9, pp. 5625–5639, Sep. 2012.

[14] C. Li, S. Weber, and J. M. Walsh, "On multi-source networks: Enumeration, rate region computation, and hierarchy," *IEEE Trans. Inf. Theory*, vol. 63, no. 11, pp. 7283–7303, Nov. 2017.

[15] C. Li, "On rate region of caching problems with non-uniform file and cache sizes," *IEEE Commun. Lett.*, vol. 21, no. 2, pp. 238–241, Feb. 2017.

[16] C. Li, S. Weber, and J. M. Walsh, "Multilevel diversity coding systems: Rate regions, codes, computation, & forbidden minors," *IEEE Trans. Inf. Theory*, vol. 63, no. 1, pp. 230–251, Jan. 2017.

[17] Z. Zhang and R. W. Yeung, "A non-Shannon-type conditional inequality of information quantities," *IEEE Trans. Inf. Theory*, vol. 43, no. 6, pp. 1982–1986, Nov. 1997.

[18] Z. Zhang and R. W. Yeung, "On characterization of entropy function via information inequalities," *IEEE Trans. Inf. Theory*, vol. 44, no. 4, pp. 1440–1452, Jul. 1998.

[19] F. Matus, "Infinitely many information inequalities," in *Proc. IEEE Int. Symp. Inf. Theory*, Jun. 2007, pp. 41–44.

[20] R. Dougherty, C. Freiling, and K. Zeger, "Six new non-Shannon information inequalities," in *Proc. IEEE Int. Symp. Inf. Theory*, Jul. 2006, pp. 233–236.

[21] R. Lněnička, "On the tightness of the Zhang-Yeung inequality for Gaussian vectors," *Commun. Inf. Syst.*, vol. 3, no. 1, pp. 41–46, 2003.

[22] K. Makarychev, Y. Makarychev, A. Romashchenko, and N. Vereshchagin, "A new class of non-Shannon-type inequalities for entropies," *Commun. Inf. Syst.*, vol. 2, no. 2, pp. 147–166, 2002.

[23] S. Prajna, A. Papachristodoulou, and P. A. Parrilo, "Introducing SOSTOOLS: A general purpose sum of squares programming solver," in *Proc. 41st IEEE Conf. Decis. Control*, vol. 1, Dec. 2002, pp. 741–746.

[24] R. W. Yeung, *Information Theory and Network Coding*. New York, NY, USA: Springer, 2008.

[25] D. Bertsimas and J. N. Tsitsiklis, *Introduction to Linear Optimization*. Belmont, MA, USA: Athena Scientific, 1997.

[26] M. Wang, C. W. Tan, W. Xu, and A. Tang, "Cost of not splitting in routing: Characterization and estimation," *IEEE/ACM Trans. Netw.*, vol. 19, no. 6, pp. 1849–1859, Dec. 2011.

[27] D. Avis and K. Fukuda, "A pivoting algorithm for convex hulls and vertex enumeration of arrangements and polyhedra," *Discrete Comput. Geometry*, vol. 8, no. 3, pp. 295–313, Sep. 1992.

[28] E. J. Candes and T. Tao, "Decoding by linear programming," *IEEE Trans. Inf. Theory*, vol. 51, no. 12, pp. 4203–4215, Dec. 2005.

[29] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[30] D. L. Donoho and Y. Tsaig, "Fast solution of $\ell_1$-norm minimization problems when the solution may be sparse," *IEEE Trans. Inf. Theory*, vol. 54, no. 11, pp. 4789–4812, Nov. 2008.

[31] J. M. Walsh and S. Weber, "A recursive construction of the set of binary entropy vectors and related algorithmic inner bounds for the entropy region," *IEEE Trans. Inf. Theory*, vol. 57, no. 10, pp. 6356–6363, Oct. 2011.

[32] S. Alam, S. Thakor, and S. Abbas, "On enumerating distributions for associated vectors in the entropy space," in *Proc. Int. Symp. Inf. Theory Its Appl. (ISITA)*, Oct. 2018, pp. 65–69.

[33] D. A. Spielman and S.-H. Teng, "Smoothed analysis of algorithms: Why the simplex algorithm usually takes polynomial time," *J. ACM*, vol. 51, no. 3, pp. 385–463, 2004.

[34] R. E. Bixby and M. J. Saltzman, "Recovering an optimal LP basis from an interior point solution," *Oper. Res. Lett.*, vol. 15, no. 4, pp. 169–178, May 1994.

[35] P. E. Gill, W. Murray, M. A. Saunders, J. A. Tomlin, and M. H. Wright, "On projected newton barrier methods for linear programming and an equivalence to Karmarkar's projective method," *Math. Program.*, vol. 36, no. 2, pp. 183–209, 1986.

[36] R. E. Marsten, M. J. Saltzman, D. F. Shanno, G. S. Pierce, and J. F. Ballintijn, "Implementation of a dual affine interior point algorithm for linear programming," *ORSA J. Comput.*, vol. 1, no. 4, pp. 287–297, Nov. 1989.

[37] S. Thakor, T. Chan, and A. Grant, "A minimal set of Shannon-type inequalities for functional dependence structures," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2017, pp. 679–683.

[38] T. Chan, S. Thakor, and A. Grant, "Minimal characterization of Shannon-type inequalities under functional dependence and full conditional independence structures," *IEEE Trans. Inf. Theory*, vol. 65, no. 7, pp. 4041–4051, Jul. 2019.

[39] S. Boyd, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Found. Trends Mach. Learn.*, vol. 3, no. 1, pp. 1–122, 2011.

[40] N. Parikh and S. Boyd, "Proximal algorithms," *Found. Trends Optim.*, vol. 1, no. 3, pp. 127–239, 2014.

[41] R. Van der Merwe and E. A. Wan, "The square-root unscented Kalman filter for state and parameter-estimation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 6, May 2001, pp. 3461–3464.

[42] Gurobi Optimization, LLC. (2018). *Gurobi Optimizer Reference Manual*. [Online]. Available: http://www.gurobi.com

[43] J. Nickolls, I. Buck, M. Garland, and K. Skadron, "Scalable parallel programming with CUDA," in *Proc. ACM SIGGRAPH Classes*. New York, NY, USA: ACM, 2008, p. 16.

[44] M. Pawłowski, T. Paterek, D. Kaszlikowski, V. Scarani, A. Winter, and M. Żukowski, "Information causality as a physical principle," *Nature*, vol. 461, no. 7267, pp. 1101–1104, Oct. 2009.

[45] R. Chaves, C. Majenz, and D. Gross, "Information–theoretic implications of quantum causal structures," *Nature Commun.*, vol. 6, no. 1, p. 5766, May 2015.

[46] K. Benidis, Y. Feng, and D. P. Palomar, "Sparse portfolios for high-dimensional financial index tracking," *IEEE Trans. Signal Process.*, vol. 66, no. 1, pp. 155–170, Jan. 2018.

[47] C. W. Tan and L. Ling, "Automated reasoning by convex optimization: Proof simplicity, duality and sparsity," in *Proc. 54th Annu. Conf. Inf. Sci. Syst. (CISS)*, 2020.

**Siu-Wai Ho** (Senior Member, IEEE) received the B.Eng., M.Phil., and Ph.D. degrees in information engineering from The Chinese University of Hong Kong, Hong Kong, in 2000, 2003, and 2006, respectively.

From 2006 to 2008, he was a Post-Doctoral Research Fellow of the Department of Electrical Engineering, Princeton University, Princeton, NJ, USA. He joined the Institute for Telecommunications Research, University of South Australia (UniSA), Adelaide, SA, Australia, in 2009, and has been a Senior Research Fellow since 2013. Since 2018, he has been a Senior Research Fellow of the Teletraffic Research Centre, University of Adelaide. His current research interests include visible light communications, indoor positioning, Shannon theory, information-theoretic security, and biometric security systems. He received the Croucher Foundation Fellowship from 2006 to 2008, the 2008 Young Scientist Award from the Hong Kong Institution of Science, the UniSA Research SA Fellowship from 2010 to 2013, and the Australian Post-Doctoral Fellowship of the Australian Research Council from 2010 to 2013. His project received the 2016 National Award—Consumer Category from the Australian Information Industry Association. He was a corecipient of the Best Paper Award from the IEEE/IET International Symposium on Communication Systems, Networks, and Digital Signal Processing 2016 and the Best Student Paper Award from the 2016 Australian Communication Theory Workshop. With his Ph.D. students, his project received an honorary mention from the 2015 ComSoc Student Competition Communications Technology Changing the World organized by the IEEE Communications Society.

**Lin Ling** received the B.Sc. degree in applied physics from the City University of Hong Kong in 2017, where he is currently pursuing the Ph.D. degree with the Department of Computer Science. He is currently a Visiting Student Collaborator with the Department of Electrical Engineering, Princeton University. His research interests include convex optimization and distributed algorithms, as well as their applications in graph analytics, machine learning, and large-scale education.

**Chee Wei Tan** (Senior Member, IEEE) received the M.A. and Ph.D. degrees in electrical engineering from Princeton University. He was a Post-Doctoral Scholar with the California Institute of Technology (Caltech). He is currently a Professor in computer science with the City University of Hong Kong. He was a Senior Fellow of the Institute for Pure and Applied Mathematics for the program on Science at Extreme Scales: Where Big Data Meets Large-Scale Computing and a Visiting Faculty with Qualcomm Research and Development and Tencent AI Lab. His research interests include artificial intelligence, networks and graph analytics, scientific machine learning, and convex optimization theory. He was a recipient of the Princeton University Wu Prize for Excellence, the Google Faculty Award, and the IEEE Information Theory Society Hong Kong Chapter of the Year Award. He currently serves as an Editor for the IEEE/ACM TRANSACTIONS ON NETWORKING.

**Raymond W. Yeung** (Fellow, IEEE) was born in Hong Kong, in June 3, 1962. He received the B.S., M.Eng., and Ph.D. degrees in electrical engineering from Cornell University, Ithaca, NY, USA, in 1984, 1985, and 1988, respectively.

He was on leave at Ecole Nationale Supérieure des Télécommunications, Paris, France, in 1986. He was a Member of Technical Staff of AT&T Bell Laboratories from 1988 to 1991. Since 1991, he has been with The Chinese University of Hong Kong, where he is currently a Choh-Ming Li Professor of information engineering and the Co-Director of the Institute of Network Coding. He has held visiting positions at Cornell University, Nankai University, the University of Bielefeld, the University of Copenhagen, the Tokyo Institute of Technology, the Munich University of Technology, and Columbia University. He was a consultant of a project with the Jet Propulsion Laboratory, Pasadena, CA, for salvaging the malfunctioning Galileo spacecraft and a consultant of NEC, USA. His 25-bit synchronization marker was used onboard, the Galileo Spacecraft, for image synchronization. He has authored textbooks *A First Course in Information Theory* (Kluwer Academic/Plenum 2002) and its revision *Information Theory and Network Coding* (Springer, 2008), which have been adopted by over 100 institutions around the world. This book has also been published in Chinese (Higher Education Press 2011, translation by N. Cai *et al.*). He has also coauthored with S. Yang the monograph *BATS Codes: Theory and Applications* (Morgan & Claypool Publishers, 2017). In 2014, he gave the first MOOC on information theory that reached over 25 000 students. His research interests include information theory and network coding.

Dr. Yeung is a fellow of the Hong Kong Academy of Engineering Sciences and the Hong Kong Institution of Engineers. He was a member of the Board of Governors of the IEEE Information Theory Society from 1999 to 2001. He has served on the committees of a number of information theory symposiums and workshops. He was the General Chair of the First and the Fourth Workshops on Network, Coding, and Applications (NetCod 2005 and 2008), a Technical Co-Chair of the 2006 IEEE International Symposium on Information Theory, the 2006 IEEE Information Theory Workshop (Chengdu, China), and the General Co-Chair of the 2015 IEEE International Symposium on Information Theory. From 2011 to 2012, he serves as a Distinguished Lecturer of the IEEE Information Theory Society. He was a recipient of the Croucher Foundation Senior Research Fellowship from 2000 to 2001, the Best Paper Award (Communication Theory) from the 2004 International Conference on Communications, Circuits, and Systems, the 2005 IEEE Information Theory Society Paper Award, the Friedrich Wilhelm Bessel Research Award of the Alexander von Humboldt Foundation in 2007, the 2016 IEEE Eric E. Sumner Award (for pioneering contributions to the field of network coding), and the 2018 ACM SIGMOBILE Test-of-Time Paper Award. In 2015, he was named (together with Z. Zhang) an Outstanding Overseas Chinese Information Theorist by the China Information Theory Society. In 2019, his team won a Gold Medal with Congratulations of the Jury at the 47th International Exhibition of Inventions of Geneva for their invention BATS: Enabling the Nervous System of Smart Cities. He currently serves as an Editor-at-Large of *Communications in Information and Systems*, an Editor of *Foundation and Trends in Communications and Information Theory*, and *Foundation and Trends in Networking*. He was an Associate Editor of Shannon Theory of the IEEE TRANSACTIONS ON INFORMATION THEORY from 2003 to 2005.