# To Prove or to Disprove: Information Inequalities Prover, Automated Reasoning by Convex Optimization & Software-as-a-Service

Chee Wei Tan

In collaboration with Alex Lin Ling, Siu Wai Ho and Raymond W. Yeung
Institute of Pure and Applied Mathematics, December 5th 2018

# Table of contents

# Information Theoretic Inequalities

Shannon's Information Measures:

- entropy: $H(A)$

# Information Theoretic Inequalities

Shannon's Information Measures:

- entropy: $H(A)$
- mutual information: $I(A; B)$

# Information Theoretic Inequalities

Shannon's Information Measures:

- entropy: $H(A)$
- mutual information: $I(A; B)$
- conditional entropy: $H(A|B)$
- conditional mutual information: $I(A; B|C)$

# Information Theoretic Inequalities

Shannon's Information Measures:

- entropy: $H(A)$
- mutual information: $I(A; B)$
- conditional entropy: $H(A|B)$
- conditional mutual information: $I(A; B|C)$

Information expression: linear combination of information measures involving finite number of random variables.

# Information Theoretic Inequalities

Shannon's Information Measures:

- entropy: $H(A)$
- mutual information: $I(A; B)$
- conditional entropy: $H(A|B)$
- conditional mutual information: $I(A; B|C)$

Information expression: linear combination of information measures involving finite number of random variables.

Information inequality: $f \geq c$, where $f$ is an information expression and $c$ is a constant.

# Information Theoretic Inequalities

Some information inequalities are always true, e.g.,

$$H(A,B,C) \geq I(A;B)$$

while others may not, e.g.,

$$H(A,B,C) \geq I(A;B) + I(B;C)$$

# Information Theoretic Inequalities

Some information inequalities are always true, e.g.,

$$H(A, B, C) \geq I(A; B)$$

while others may not, e.g.,

$$H(A, B, C) \geq I(A; B) + I(B; C)$$

Given an information theoretic inequality, is there an algorithm to prove or disprove it automatically?

# The ITIP Framework: Information-Theoretic Inequality Prover by Linear Programming

# Information Expressions in Canonical Form

Any information expression can be represented as a linear combination of entropies and joint-entropies. e.g.:

$$
\begin{aligned}
H(A|B) &= H(A,B) - H(B) \\
I(A;B) &= H(A) + H(B) - H(A,B) \\
I(A;B|C) &= H(A,C) + H(B,C) - H(A,B,C) - H(C)
\end{aligned}
$$

For $n$ random variables, there are $k = 2^n - 1$ entropies/joint-entropies.

# Information Expressions in Canonical Form

Any information expression can be represented as a linear combination of entropies and joint-entropies. e.g.:

$$\begin{aligned} H(A|B) &= H(A,B) - H(B) \\ I(A;B) &= H(A) + H(B) - H(A,B) \\ I(A;B|C) &= H(A,C) + H(B,C) - H(A,B,C) - H(C) \end{aligned}$$

For $n$ random variables, there are $k = 2^n - 1$ entropies/joint-entropies.

e.g., for $n = 3$, we have:

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

# Information Expressions in Canonical Form

This inspires us to express information inequalities in form of $b^T h \geq 0$, where $h \in \mathbb{R}^k$. e.g.

$$H(A|B) \geq I(A;B) \Longrightarrow 2H(A,B) - 2H(B) - H(A) \geq 0$$

| H(A) | H(B) | H(AB) |
|---|---|---|
| -1 | -2 | 2 |

$$\Longrightarrow b^T h \geq 0,$$

where

$$b = \begin{bmatrix} -1 & -2 & 2 \end{bmatrix}^T$$

# Elemental Inequalities

A valid $h$ has to ensure nonnegativity of all information measures.

e.g., for $n = 2$, we must have

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

# Elemental Inequalities

A valid $h$ has to ensure nonnegativity of all information measures.

e.g., for $n = 2$, we must have

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

This is equivalent to $Dh \geq 0$, where

$$D = \begin{bmatrix} 0 & -1 & 1 \\ -1 & 0 & 1 \\ 1 & 1 & -1 \end{bmatrix}$$

# Elemental Inequalities

For $n$ random variables, there are

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

elemental inequalities to satisfy, thus $D \in \mathbb{R}^{m \times n}$

# Linear Programming Formulation

From the above discussion, we notice that to prove or disprove a given information inequality, we can calculate the lower bound of $b^T h$. This can be formulated into a simple LP:

$$\begin{aligned} \min \quad & p = b^T h \\ \text{s.t.} \quad & Dh \geq 0 \end{aligned}$$

If $p^* \geq 0$, we know the corresponding inequality is always true, otherwise we conclude that it does not always hold.

# Proof/Disproof Construction

Wait a minute... What this does is just verification, not proving. What we want is human-readable proof/disproof.

# Proof/Disproof Construction

Consider the dual problem:

$$\begin{aligned} \max \quad & 0 \\ \text{s.t.} \quad & b - D^T\lambda = 0 \\ & \lambda \geq 0 \end{aligned}$$

Assume the problem is feasible. Let the optimal solution be $\lambda^*$, we know $b = D^T\lambda^*$

# Proof/Disproof Construction

Consider the dual problem:

$$\begin{aligned} \max \quad & 0 \\ \text{s.t.} \quad & b - D^T\lambda = 0 \\ & \lambda \geq 0 \end{aligned}$$

Assume the problem is feasible. Let the optimal solution be $\lambda^*$, we know $b = D^T\lambda^*$

Notice that, for **ANY** feasible $h$, we have

$$b^T h = {\lambda^*}^T D h \geq 0$$

# Proof/Disproof Construction

Consider the dual problem:

$$\begin{aligned} \max \quad & 0 \\ \text{s.t.} \quad & b - D^T\lambda = 0 \\ & \lambda \geq 0 \end{aligned}$$

Assume the problem is feasible. Let the optimal solution be $\lambda^*$, we know $b = D^T\lambda^*$

Notice that, for **ANY** feasible $h$, we have

$$b^T h = \lambda^{*T} D h \geq 0$$

This is our proof!

# Toy Example

Let's prove $2H(A,B) >= H(A) + H(B) \implies b = \begin{bmatrix} -1 & -1 & 2 \end{bmatrix}^T$,

$$D = \begin{bmatrix} 0 & -1 & 1 \\ -1 & 0 & 1 \\ 1 & 1 & -1 \end{bmatrix} \quad \begin{array}{c} H(A|B) = H(A,B) - H(B) \geq 0 \\ H(B|A) = H(A,B) - H(A) \geq 0 \\ I(A;B) = H(A) + H(B) - H(A,B) \geq 0 \end{array}$$

# Toy Example

Let's prove $2H(A,B) >= H(A) + H(B) \Longrightarrow b = \begin{bmatrix} -1 & -1 & 2 \end{bmatrix}^T$,

$$D = \begin{bmatrix} 0 & -1 & 1 \\ -1 & 0 & 1 \\ 1 & 1 & -1 \end{bmatrix} \quad \begin{matrix} H(A|B) = H(A,B) - H(B) \geq 0 \\ H(B|A) = H(A,B) - H(A) \geq 0 \\ I(A;B) = H(A) + H(B) - H(A,B) \geq 0 \end{matrix}$$

Solve that $\lambda^* = \begin{bmatrix} 1 & 1 & 0 \end{bmatrix}^T$

# Toy Example

Let's prove $2H(A,B) >= H(A) + H(B) \implies b = \begin{bmatrix} -1 & -1 & 2 \end{bmatrix}^T$,

$$D = \begin{bmatrix} 0 & -1 & 1 \\ -1 & 0 & 1 \\ 1 & 1 & -1 \end{bmatrix} \quad \begin{aligned} H(A|B) = H(A,B) - H(B) \geq 0 \\ H(B|A) = H(A,B) - H(A) \geq 0 \\ I(A;B) = H(A) + H(B) - H(A,B) \geq 0 \end{aligned}$$

Solve that $\lambda^* = \begin{bmatrix} 1 & 1 & 0 \end{bmatrix}^T$

Proof:

$$\begin{aligned} & 2H(A,B) - H(A) - H(B) \\ &= (H(A,B) - H(B)) + (H(A,B) - H(A)) \\ &= H(A|B) + H(B|A) \geq 0 \end{aligned}$$

# Disproof Construction

The primal would be unbounded below if the inequality is not provable.

# Disproof Construction

The primal would be unbounded below if the inequality is not provable.

There exists a subset of entries of $h^*$ that's $\infty \implies$ Add an upper bound to $h$

e.g. let $H(X_1, X_2, \cdots, X_n) = 1$.

$$\begin{aligned} \min \quad & b^T h \\ \text{s.t.} \quad & Dh \geq 0 \\ & e^T h = 1 \end{aligned}$$

, where $e = \left[0, 0, \cdots, 1\right]^T$

$$\begin{aligned} \max \quad & -\mu \\ \text{s.t.} \quad & b - D^T \lambda + \mu e = 0 \\ & \lambda \geq 0 \end{aligned}$$

# The Computational Challenge

# $D$ Matrix Size

$D \in \mathbb{R}^{m \times k}$, where $m = n + \binom{n}{2} 2^{n-2}$ and $k = 2^n - 1$

# $D$ Matrix Size

$D \in \mathbb{R}^{m \times k}$, where $m = n + \binom{n}{2} 2^{n-2}$ and $k = 2^n - 1$

The bad news: The size of $D$ grows exponentially!

# $D$ Matrix Size

$D \in \mathbb{R}^{m \times k}$, where $m = n + \binom{n}{2} 2^{n-2}$ and $k = 2^n - 1$

The bad news: The size of $D$ grows exponentially!

The good news: $D$ is very sparse

# $D$ Matrix Size

$D \in \mathbb{R}^{m \times k}$, where $m = n + \binom{n}{2} 2^{n-2}$ and $k = 2^n - 1$

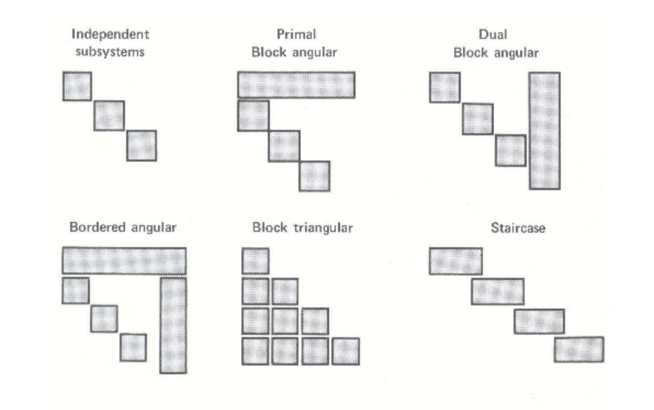
The bad news: The size of $D$ grows exponentially!

The good news: $D$ is very sparse

| n | m | k | sparsity |
|---|---|---|---|
| 2 | 3 | 3 | 0.222222 |
| 5 | 85 | 31 | 0.878558 |
| 10 | 11530 | 1023 | 0.996095 |
| 15 | 860175 | 32767 | 0.999878 |
| 20 | 49807380 | 1048575 | 0.999996 |

# $D$ Matrix Spatial Structure

One of the standard tricks to deal with large-scale LPs is to look for spatial structures in the constraint matrix ($D$ in our case) and decompose the LP into a number of smaller LPs.



# $D$ Matrix Spatial Structure

Unfortunately this cannot be done for our $D$ matrix.

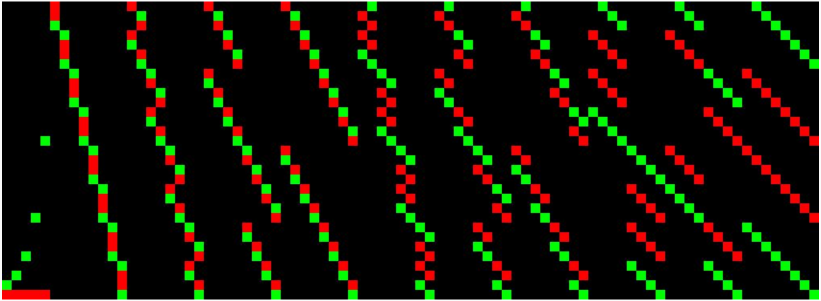


**Figure 1:** Sparsity map for $D^T$ when $n = 5$

No spatial structure that we can utilize :(

# Degeneracy

It's easy to see that in the dual problem

$$\begin{aligned} \max \quad & 0 \\ \text{s.t.} \quad & b - D^T\lambda = 0 \\ & \lambda \geq 0, \end{aligned}$$

there's an infinite number of optimal solution $\lambda^*$, which means the problem is highly degenerate.

# Degeneracy

It's easy to see that in the dual problem

$$\begin{aligned} \max \quad & 0 \\ \text{s.t.} \quad & b - D^T\lambda = 0 \\ & \lambda \geq 0, \end{aligned}$$

there's an infinite number of optimal solution $\lambda^*$, which means the problem is highly degenerate.

The simplex method would struggle on degenerate problems.

# Degeneracy

It's easy to see that in the dual problem

$$\begin{aligned} \max \quad & 0 \\ \text{s.t.} \quad & b - D^T\lambda = 0 \\ & \lambda \geq 0, \end{aligned}$$

there's an infinite number of optimal solution $\lambda^*$, which means the problem is highly degenerate.

The simplex method would struggle on degenerate problems.

Can we use numerical algorithms (e.g. interior-point method)? Yes, but there's a catch...

# An ADMM-Based Scalable Solution

# Generic ADMM

Consider the alternating direction method of multipliers (ADMM) to solve the following optimization problem:

$$\begin{aligned} \min \quad & f(x) + g(y) \\ \text{s.t.} \quad & Ax + By = c, \end{aligned}$$

with $\rho$-augmented Lagrangian:

$$L_\rho = f(x) + g(y) + \lambda^T(Ax + By - c) + \frac{\rho}{2}||Ax + By - c||^2,$$

Consider the alternating direction method of multipliers (ADMM) to solve the following optimization problem:

$$\begin{aligned} \min \quad & f(x) + g(y) \\ \text{s.t.} \quad & Ax + By = c, \end{aligned}$$

with $\rho$-augmented Lagrangian:

$$L_\rho = f(x) + g(y) + \lambda^T(Ax + By - c) + \frac{\rho}{2}||Ax + By - c||^2,$$

a generic ADMM algorithm is given as follows:

**ALGORITHM 2:** Generic ADMM Algorithm

**repeat**

1. $x$-update: $x^{k+1} = \arg\min\{L_\rho(x, y^k, \lambda^k)\}$
2. $y$-update: $y^{k+1} = \arg\min\{L_\rho(x^{k+1}, y, \lambda^k)\}$
3. $\lambda$-update: $\lambda^{k+1} = \lambda^k + \rho(Ax^{k+1} + By^{k+1} - c)$

**until** *Stopping criteria is met*;

# Problem Reformulation

Consider the following reformulation:

$$\begin{aligned} \min \quad & b^T h \\ \text{s.t.} \quad & 0 \leq Ah \leq 1, \end{aligned}$$

where $A = \begin{bmatrix} D \\ I \end{bmatrix}$.

# Problem Reformulation

Consider the following reformulation:

$$\begin{aligned} \min \quad & b^T h \\ \text{s.t.} \quad & 0 \leq Ah \leq 1, \end{aligned}$$

where $A = \begin{bmatrix} D \\ I \end{bmatrix}$.

1. If the inequality is provable, the extra constraint $Ah \leq 1$ would be redundant. In this case, the reformulated problem is equivalent to the original one in both primal and dual.

# Problem Reformulation

Consider the following reformulation:

$$\begin{aligned} \min \quad & b^T h \\ \text{s.t.} \quad & 0 \leq Ah \leq 1, \end{aligned}$$

where $A = \begin{bmatrix} D \\ I \end{bmatrix}$.

1. If the inequality is provable, the extra constraint $Ah \leq 1$ would be redundant. In this case, the reformulated problem is equivalent to the original one in both primal and dual.
2. If the inequality is not provable, we are still upper-bounding the entries of $h$ as the original problem does, and we can still construct disproof using the optimal dual values.

# Problem Reformulation

Consider the following reformulation:

$$\begin{aligned} \min \quad & b^T h \\ \text{s.t.} \quad & 0 \leq Ah \leq 1, \end{aligned}$$

where $A = \begin{bmatrix} D \\ I \end{bmatrix}$.

1. If the inequality is provable, the extra constraint $Ah \leq 1$ would be redundant. In this case, the reformulated problem is equivalent to the original one in both primal and dual.
2. If the inequality is not provable, we are still upper-bounding the entries of $h$ as the original problem does, and we can still construct disproof using the optimal dual values.

We can just solve this one problem!

# Problem Reformulation

Now we reformulate the problem to ADMM form by adding slack variables $u$ and $v$:

$$\begin{aligned} \min \quad & b^T h \\ \text{s.t.} \quad & Bh + z = c, \end{aligned}$$

where $B = \begin{bmatrix} A \\ A \end{bmatrix}, z = \begin{bmatrix} -u \\ v \end{bmatrix}, c = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, u, v \geq 0$

# ADMM Subproblem Computation

The $\rho$-augmented Lagrangian of the problem is:

$$L_\rho = b^T h + \lambda^T (Bh + z - c) + \frac{\rho}{2} ||Bh + z - c||^2$$

# ADMM Subproblem Computation

The $\rho$-augmented Lagrangian of the problem is:

$$L_\rho = b^T h + \lambda^T (Bh + z - c) + \frac{\rho}{2} ||Bh + z - c||^2$$

$h$-update:

$$h^{k+1} = \arg\min L_\rho(h, z^k, \lambda^k)$$

# ADMM Subproblem Computation

The $\rho$-augmented Lagrangian of the problem is:

$$L_\rho = b^T h + \lambda^T (Bh + z - c) + \frac{\rho}{2} ||Bh + z - c||^2$$

$h$-update:

$$h^{k+1} = \arg\min L_\rho(h, z^k, \lambda^k)$$

This is an unconstrained QP, and we have a closed-form solution:

$$h^{k+1} = -\frac{1}{\rho}(B^T B)^{-1}(b + B^T \lambda^k + \rho B^T z^k - \rho B^T c).$$

It's easy to prove that $(B^T B)^{-1}$ exists.

# ADMM Subproblem Computation

$z$-update:

$$z^{k+1} = \arg\min L_\rho(h^{k+1}, z, \lambda^k)$$

Recall that $z = \begin{bmatrix} -u & v \end{bmatrix}^T$, so we can split this subproblem into $u$-update and $v$-update:

$$u^{k+1} = \arg\min\{L_\rho(h^{k+1}, u, \lambda^k) | u \geq 0\}$$
$$v^{k+1} = \arg\min\{L_\rho(h^{k+1}, v, \lambda^k) | v \geq 0\}$$

# ADMM Subproblem Computation

$z$-update:

$$z^{k+1} = \arg\min L_\rho(h^{k+1}, z, \lambda^k)$$

Recall that $z = \begin{bmatrix} -u & v \end{bmatrix}^T$, so we can split this subproblem into $u$-update and $v$-update:

$$u^{k+1} = \arg\min\{L_\rho(h^{k+1}, u, \lambda^k) | u \geq 0\}$$
$$v^{k+1} = \arg\min\{L_\rho(h^{k+1}, v, \lambda^k) | v \geq 0\}$$

They are constrained QPs, but luckily their KKT systems can be solved directly, giving us closed-form solutions:

$$u^{k+1} = (Ah^{k+1} + \frac{1}{\rho}\lambda_u^k)_+$$
$$v^{k+1} = (1 - Ah^{k+1} - \frac{1}{\rho}\lambda_v^k)_+$$

# ADMM Algorithm

**ALGORITHM 3:** ITIP Algorithm

**repeat**

1. $h$-update: $h^{k+1} = -\frac{1}{\rho}(B^T B)^{-1}(b + B^T \lambda^k + \rho B^T z^k - \rho B^T c)$
2. $u$-update: $u^{k+1} = (Ah^{k+1} + \frac{1}{\rho}\lambda_u^k)_+$
3. $v$-update: $v^{k+1} = (1 - Ah^{k+1} - \frac{1}{\rho}\lambda_v^k)_+$
4. $\lambda$-update: $\lambda^{k+1} = \lambda^k + \rho(Bh^{k+1} + z^{k+1} - c)$

**until** *Stopping criterion is met*;

# Why This Algorithm?

The advantages:

1. Have simple closed-form solution for every subproblem (linear algebraic structures enable scalable and distributed computation, CPU-GPU configuration, cloud computing)

# Why This Algorithm?

The advantages:

1. Have simple closed-form solution for every subproblem (linear algebraic structures enable scalable and distributed computation, CPU-GPU configuration, cloud computing)
2. Not simplex-based, so degeneracy is not a problem

# Why This Algorithm?

The advantages:

1. Have simple closed-form solution for every subproblem (linear algebraic structures enable scalable and distributed computation, CPU-GPU configuration, cloud computing)
2. Not simplex-based, so degeneracy is not a problem
3. Make good use of sparsity of $D$

# Why This Algorithm?

The advantages:

1. Have simple closed-form solution for every subproblem (linear algebraic structures enable scalable and distributed computation, CPU-GPU configuration, cloud computing)
2. Not simplex-based, so degeneracy is not a problem
3. Make good use of sparsity of $D$
4. Offload computational burden to inverting $B^T B$, which can be pre-computed and cached

# Why This Algorithm?

The advantages:

1. Have simple closed-form solution for every subproblem (linear algebraic structures enable scalable and distributed computation, CPU-GPU configuration, cloud computing)
2. Not simplex-based, so degeneracy is not a problem
3. Make good use of sparsity of $D$
4. Offload computational burden to inverting $B^T B$, which can be pre-computed and cached

The disadvantage:

- Need to “crossover” to obtain “elegent” proofs/disproofs

# What is in a Proof?

Let's prove $H(XYZ) - H(X|YZ) - H(Y|XZ) - H(Z|XY) \geq 0$

# What is in a Proof?

Let's prove $H(XYZ) - H(X|YZ) - H(Y|XZ) - H(Z|XY) \geq 0$

$$
\begin{aligned}
& H(XYZ) - H(X|YZ) - H(Y|XZ) - H(Z|XY) \\
= & I(X;Y) + I(X;Z|Y) + I(Y;Z|X) \qquad (1) \\
= & I(X;Z) + I(X;Y|Z) + I(Y;Z|X) \qquad (2) \\
= & I(Y;Z) + I(X;Z|Y) + I(X;Y|Z). \qquad (3)
\end{aligned}
$$

# What is in a Proof?

Let's prove $H(XYZ) - H(X|YZ) - H(Y|XZ) - H(Z|XY) \geq 0$

$$\begin{aligned} & H(XYZ) - H(X|YZ) - H(Y|XZ) - H(Z|XY) \\ =& I(X;Y) + I(X;Z|Y) + I(Y;Z|X) \quad (1) \\ =& I(X;Z) + I(X;Y|Z) + I(Y;Z|X) \quad (2) \\ =& I(Y;Z) + I(X;Z|Y) + I(X;Y|Z). \quad (3) \end{aligned}$$

In fact,

$$\begin{aligned} & H(XYZ) - H(X|YZ) - H(Y|XZ) - H(Z|XY) \\ =& 0.8(I(X;Y) + I(X;Z|Y) + I(Y;Z|X)) + \\ & 0.1(I(X;Z) + I(X;Y|Z) + I(Y;Z|X)) + \\ & 0.1(I(Y;Z) + I(X;Z|Y) + I(X;Y|Z)) \quad (4) \end{aligned}$$

is also true.

# What is in a Proof?

Let's prove $H(XYZ) - H(X|YZ) - H(Y|XZ) - H(Z|XY) \geq 0$

$$
\begin{aligned}
& H(XYZ) - H(X|YZ) - H(Y|XZ) - H(Z|XY) \\
=& I(X;Y) + I(X;Z|Y) + I(Y;Z|X) \qquad (1) \\
=& I(X;Z) + I(X;Y|Z) + I(Y;Z|X) \qquad (2) \\
=& I(Y;Z) + I(X;Z|Y) + I(X;Y|Z). \qquad (3)
\end{aligned}
$$

In fact,

$$
\begin{aligned}
& H(XYZ) - H(X|YZ) - H(Y|XZ) - H(Z|XY) \\
=& 0.8(I(X;Y) + I(X;Z|Y) + I(Y;Z|X)) + \\
& 0.1(I(X;Z) + I(X;Y|Z) + I(Y;Z|X)) + \\
& 0.1(I(Y;Z) + I(X;Z|Y) + I(X;Y|Z)) \qquad (4)
\end{aligned}
$$

is also true. Obviously, the proof in (4) is unnecessarily complicated and not succinctly elegant.

# What is in a Proof?

Let's prove $H(XYZ) - H(X|YZ) - H(Y|XZ) - H(Z|XY) \geq 0$

$$
\begin{aligned}
& H(XYZ) - H(X|YZ) - H(Y|XZ) - H(Z|XY) \\
=& I(X;Y) + I(X;Z|Y) + I(Y;Z|X) && (1) \\
=& I(X;Z) + I(X;Y|Z) + I(Y;Z|X) && (2) \\
=& I(Y;Z) + I(X;Z|Y) + I(X;Y|Z). && (3)
\end{aligned}
$$

In fact,

$$
\begin{aligned}
& H(XYZ) - H(X|YZ) - H(Y|XZ) - H(Z|XY) \\
=& 0.8(I(X;Y) + I(X;Z|Y) + I(Y;Z|X)) + \\
& 0.1(I(X;Z) + I(X;Y|Z) + I(Y;Z|X)) + \\
& 0.1(I(Y;Z) + I(X;Z|Y) + I(X;Y|Z)) && (4)
\end{aligned}
$$

is also true. Obviously, the proof in (4) is unnecessarily complicated and not succinctly elegant.
"Beauty is the first test: there is no permanent place in this world for ugly mathematics." - G. H. Hardy

# What is an “Elegant” Proof

1. ”Everything should be made as simple as possible, but not simpler” - Albert Einstein

# What is an “Elegant” Proof

1. ”Everything should be made as simple as possible, but not simpler” - Albert Einstein
2. ”It is my experience that proofs involving matrices can be shortened by 50% if one throws the matrices out.” - Emil Artin

# What is an “Elegant” Proof

1. ”Everything should be made as simple as possible, but not simpler” - Albert Einstein
2. ”It is my experience that proofs involving matrices can be shortened by 50% if one throws the matrices out.” - Emil Artin
3. ”This one’s from the Book!” - Paul Erdős
   *referring to a mythical ”book” in which God wrote the proofs for all theorems*. ”I know it when I see it.”

# What is an “Elegant” Proof

1. ”Everything should be made as simple as possible, but not simpler” - Albert Einstein
2. ”It is my experience that proofs involving matrices can be shortened by 50% if one throws the matrices out.” - Emil Artin
3. ”This one’s from the Book!” - Paul Erdős
   *referring to a mythical ”book” in which God wrote the proofs for all theorems.* ”I know it when I see it.”
4. ”The proof is trivial and is left as an exercise.” - Annoymous

# What is an “Elegant” Proof

1. ”Everything should be made as simple as possible, but not simpler” - Albert Einstein
2. ”It is my experience that proofs involving matrices can be shortened by 50% if one throws the matrices out.” - Emil Artin
3. ”This one’s from the Book!” - Paul Erdős
   *referring to a mythical ”book” in which God wrote the proofs for all theorems*. ”I know it when I see it.”
4. ”The proof is trivial and is left as an exercise.” - Annoymous

The beauty of mathematics contains a few parameters: It is short.

# What is an “Elegant” Proof

1. ”Everything should be made as simple as possible, but not simpler” - Albert Einstein
2. ”It is my experience that proofs involving matrices can be shortened by 50% if one throws the matrices out.” - Emil Artin
3. ”This one’s from the Book!” - Paul Erdős
   *referring to a mythical ”book” in which God wrote the proofs for all theorems*. ”I know it when I see it.”
4. ”The proof is trivial and is left as an exercise.” - Annoymous

The beauty of mathematics contains a few parameters: It is short. It is unexpected.

# What is an “Elegant” Proof

1. ”Everything should be made as simple as possible, but not simpler” - Albert Einstein
2. ”It is my experience that proofs involving matrices can be shortened by 50% if one throws the matrices out.” - Emil Artin
3. ”This one’s from the Book!” - Paul Erdős
   *referring to a mythical ”book” in which God wrote the proofs for all theorems*. ”I know it when I see it.”
4. ”The proof is trivial and is left as an exercise.” - Annoymous

The beauty of mathematics contains a few parameters: It is short. It is unexpected. It enriches people.

# What is an “Elegant” Proof

1. ”Everything should be made as simple as possible, but not simpler” - Albert Einstein
2. ”It is my experience that proofs involving matrices can be shortened by 50% if one throws the matrices out.” - Emil Artin
3. ”This one’s from the Book!” - Paul Erdős
   *referring to a mythical ”book” in which God wrote the proofs for all theorems*. ”I know it when I see it.”
4. ”The proof is trivial and is left as an exercise.” - Annoymous

The beauty of mathematics contains a few parameters: It is short. It is unexpected. It enriches people. It is applicable to other problems.
- Ron Aharoni (Mathematics poetry and beauty)

# “Shortest” Proof by Automated Reasoning

Consider the “Shortest” Proof problem:

$$\begin{aligned} \min \quad & \|\lambda\|_0 \\ \text{s.t.} \quad & b - D^T\lambda = 0 \\ & \lambda \geq 0 \end{aligned}$$

# “Shortest” Proof by Automated Reasoning

Consider the “Shortest” Proof problem:

$$\begin{aligned} \min \quad & \|\lambda\|_0 \\ \text{s.t.} \quad & b - D^T\lambda = 0 \\ & \lambda \geq 0 \end{aligned}$$

Automated reasoning to generate short computer proofs via recent convex relaxation ideas for sparse recovery in compressed sensing:

$$\begin{aligned} \min \quad & \|\lambda\|_1 \\ \text{s.t.} \quad & b - D^T\lambda = 0 \\ & \lambda \geq 0 \end{aligned}$$

This is a linear program that often yields the shortest proof.

# A Scalable Software-as-a-Service

# The ITIP Web Service (WiP)

`https://itip.algebragame.app`

# Q&A

# References i


S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein, et al.
**Distributed optimization and statistical learning via the alternating direction method of multipliers.**
*Foundations and Trends® in Machine learning*, 3(1):1–122, 2011.

S.-W. Ho, C. W. Tan, and R. W. Yeung.
**Proving and disproving information inequalities.**
In *2014 IEEE International Symposium on Information Theory (ISIT)*, pages 2814–2818. Citeseer, 2014.

R. W. Yeung.
***Information theory and network coding.***
Springer Science & Business Media, 2008.