Class Road Map - IT-I

- L19 (1/6): Overview, Communications, Gaussian Channel
- L20 (1/8): Gaussian Channel, band limitation, parallel channels, optimization and duality
- L21 (1/13): parallel channels, colored noise, feedback, matrix inequalities
- L22 (1/15): matrix inequalities, rate distortion.
  - (1/20): Monday holiday
- L23 (1/22): rate distortion for Bernoulli, Gaussian, and Multiple Gaussians with unequal noise
- L24 (1/27): main rate distortion theorem, geometry
- L25 (1/29): computing $R(D)$
- L26 (2/3): computing $R(D)$, alternating minimization
- L27 (2/5): Kolmogorov complexity
- L28 (2/10): algorithmic randomness, universal prob.,
- L29 (2/12): universal compression, LZ compression
  - (2/17): Monday, Holiday
- L30 (2/19): LZ compression, info measures
- L31 (2/24): Info measures
- L32 (2/26):
- L33 (3/3):
- L34 (3/5):
- L35 (3/10):
- L36 (3/12):

Cumulative Outstanding Reading

- Read Ch. 13 in our book (Cover & Thomas, “Information Theory”).
- Read Ch. 14 in our book (Cover & Thomas, “Information Theory”).
- Read Ch. 10 in our book (Cover & Thomas, “Information Theory”).
- Read Ch. 17 in our book (Cover & Thomas, “Information Theory”) on matrix inequalities.
- Read Ch. 9 in our book (Cover & Thomas, “Information Theory”)
- Read Ch. 5 in Boyd and Vandenberghe’s Convex Optimization book
- Read all readings assigned in EE514a, Fall 2013. (see later lectures on our previous web page (http://j.ee.washington.edu/~bilmes/classes/ee514a_fall_2013/)).
Homework

- No current outstanding HW.
No class Monday, Feb 17th.
Office hours on Mondays, 3:30-4:30.
As always, email me if you want to skype/google hangout rather than come to office hours, also at different times.
On Final Presentations

- Your task is to give a 10-15 minute presentation that summarizes 2-3 related and significant papers that come from IEEE Transactions on Information Theory (or a very related area).
- The papers must not be ones that we covered in class, although they can be related.
- You need to do the research to find the papers yourself (i.e., that is part of the assignment).
- The majority of the papers must have been published in the last 10 years (so no old or classic papers).
- Your grade will be based on how clear, understandable, and accurate your presentation is (and also milestones).
- This is a real challenge and will require significant work! Many of the papers are complex. To get a good grade, you will need to work very hard to present very complex ideas in an extremely simple yet still precise way.
- Again, don’t expect this to be easy, you might need to try a few topics until you find one that is suitable.
Final Presentation Milestones

All submissions done in PDF file format via our assignment dropbox (https://canvas.uw.edu/courses/880971/assignments)

- **Monday, Feb 24th 11:45pm, tonight:** Updated list of proposed papers decided, based on feedback. Updated writeup with more description. **Include PDFs of papers!**

- **Monday, March 3rd 11:45pm:** progress report (at most 1 page). Any background papers you needed to read to better understand your core set. Thoughts on coherent and simple unifying presentation.

- **Monday, March 10th, 11:45pm:** updated short (≤ 1 page) writeup on more details of how you will present the ideas in a simple fashion.

- **Final presentations:** Monday, March 17, 2014, 2:30–4:20pm, LOW 102. What to turn in: your slides and a short at most 4 page summary of the papers.
We’ve previously seen that Venn diagrams are a useful way to visualize the relationship between information measures (entropy, etc.)

\[
H(X, Y) = H(X) + H(Y) - H(X | Y)
\]

But is within these sets? So far, we’ve only said it is “information”

We want now to show that set theory and the relation between set theory and information theory can be made more precise in order to:

1. gain intuition
2. help prove theorems
3. lead to new (useful) information theoretic inequalities that are “non-Shannon” (i.e., not previously known).
We’ve previously seen that Venn diagrams are a useful way to visualize the relationship between information measures (entropy, etc.)

But is within these sets? So far, we’ve only said it is “information”
We’ve previously seen that Venn diagrams are a useful way to visualize the relationship between information measures (entropy, etc.)

But is within these sets? So far, we’ve only said it is “information”

We want now to show that set theory and the relation between set theory and information theory can be made more precise in order to:
We’ve previously seen that Venn diagrams are a useful way to visualize the relationship between information measures (entropy, etc.)

But is within these sets? So far, we’ve only said it is “information”

We want now to show that set theory and the relation between set theory and information theory can be made more precise in order to:

1. gain intuition
We’ve previously seen that Venn diagrams are a useful way to visualize the relationship between information measures (entropy, etc.)

But is within these sets? So far, we’ve only said it is “information”

We want now to show that set theory and the relation between set theory and information theory can be made more precise in order to:

1. gain intuition
2. help prove theorems
We’ve previously seen that Venn diagrams are a useful way to visualize the relationship between information measures (entropy, etc.)

But is within these sets? So far, we’ve only said it is “information”

We want now to show that set theory and the relation between set theory and information theory can be made more precise in order to:

1. gain intuition
2. help prove theorems
3. lead to new (useful) information theoretic inequalities that are “non-Shannon” (i.e., not previously known).
We have a set of random variables $X_1, X_2, \ldots, X_n$. 
Definitions: field, atom

- We have a set of random variables $X_1, X_2, \ldots, X_n$.
- For each random variable we associate a set $\tilde{X}_1, \tilde{X}_2, \ldots, \tilde{X}_n$. 

A field $\mathcal{F}_n$ can be generated from sets $\tilde{X}_1, \ldots, \tilde{X}_n$ by taking unions ($\bigcup$), intersections ($\bigcap$), complementation ($\tilde{X}_i^c$), and set subtractions/ differences ($\cap$) on combinations of $\tilde{X}_1, \tilde{X}_2, \ldots, \tilde{X}_n$.

An atom of $\mathcal{F}_n$ are sets of the form $\bigcap_{i=1}^{n} \tilde{Y}_i$ where $\tilde{Y}_i = \{\tilde{X}_i \text{ or } \tilde{X}_i^c\}$ (31.1)

Ex: $n = 2$, 4 such atoms.
Ex: $n = 3$, 8 atoms.
Definitions: field, atom

- We have a set of random variables $X_1, X_2, \ldots, X_n$.
- For each random variable we associate a set $\tilde{X}_1, \tilde{X}_2, \ldots, \tilde{X}_n$.
- A field $\mathcal{F}_n$ can be generated from sets $\tilde{X}_{1:n}$ by taking unions ($\bigcup$), intersections ($\bigcap$), complementation ($\tilde{X}^c$), set subtractions/difference ($\setminus$) on combinations of $\tilde{X}_1, \tilde{X}_2, \ldots, \tilde{X}_n$. 

Example:
- $n = 2$, 4 such atoms.
- $n = 3$, 8 atoms.
Definitions: field, atom

- We have a set of random variables $X_1, X_2, \ldots, X_n$.
- For each random variable we associate a set $\tilde{X}_1, \tilde{X}_2, \ldots, \tilde{X}_n$.
- A field $\mathcal{F}_n$ can be generated from sets $\tilde{X}_{1:n}$ by taking unions ($\bigcup$), intersections ($\bigcap$), complementation ($\tilde{X}^c$), set subtractions/difference ($\setminus$) on combinations of $\tilde{X}_1, \tilde{X}_2, \ldots, \tilde{X}_n$.
- An atom of $\mathcal{F}_n$ are sets of the form

$$\text{an atom} = \bigcap_{i=1}^{n} Y_i \quad \text{where} \quad Y_i = \begin{cases} \tilde{X}_i \\ \tilde{X}^c_i \end{cases}$$

(31.1)
Definitions: field, atom

- We have a set of random variables $X_1, X_2, \ldots, X_n$.
- For each random variable we associate a set $\tilde{X}_1, \tilde{X}_2, \ldots, \tilde{X}_n$.
- A field $\mathcal{F}_n$ can be generated from sets $\tilde{X}_1:n$ by taking unions ($\bigcup$), intersections ($\bigcap$), complementation ($\tilde{X}_i^c$), set subtractions/difference ($\setminus$) on combinations of $\tilde{X}_1, \tilde{X}_2, \ldots, \tilde{X}_n$.
- An atom of $\mathcal{F}_n$ are sets of the form

$$\text{an atom} = \bigcap_{i=1}^{n} Y_i \quad \text{where} \quad Y_i = \begin{cases} \tilde{X}_i \\ \tilde{X}_i^c \end{cases} \quad \text{or} \quad (31.1)$$

Ex: $n = 2$, 4 such atoms.

Ex: $n = 3$, 8 atoms
In general, there are $|\mathcal{A}| = 2^n$ atoms where $\mathcal{A}$ is the set of atoms.
In general, there are $|\mathcal{A}| = 2^n$ atoms where $\mathcal{A}$ is the set of atoms.

The atoms are disjoint: Why?
In general, there are $|\mathcal{A}| = 2^n$ atoms where $\mathcal{A}$ is the set of atoms.

The atoms are disjoint: Why? For any two distinct atoms, there is at least one factor which is a complement.
Definitions: field, atom

- In general, there are $|\mathcal{A}| = 2^n$ atoms where $\mathcal{A}$ is the set of atoms.
- The atoms are disjoint: Why? For any two distinct atoms, there is at least one factor which is a complement.
- There are $2^{2^n}$ elements in the field. Why?
Definitions: field, atom

- In general, there are $|\mathcal{A}| = 2^n$ atoms where $\mathcal{A}$ is the set of atoms.
- The atoms are disjoint: Why? For any two distinct atoms, there is at least one factor which is a complement.
- There are $2^{2^n}$ elements in the field. Why? \textit{union of disjoint atoms and each atom may be chosen or not chosen}.
Definitions: signed measure

We will be measuring these sets using a signed measure (meaning it might be positive or negative). In particular, a real-valued function $\mu$ defined on $\mathcal{F}_n$ is called a signed measure if it is set-additive, i.e., for disjoint $A$ and $B$, we have

$$\mu(A \cup B) = \mu(A) + \mu(B)$$

(31.2)
Definitions: signed measure

We will be measuring these sets using a signed measure (meaning it might be positive or negative). In particular, a real-valued function \( \mu \) defined on \( \mathcal{F}_n \) is called a signed measure if it is set-additive, i.e., for disjoint \( A \) and \( B \), we have

\[
\mu(A \cup B) = \mu(A) + \mu(B)
\]  \hspace{1cm} (31.2)

for a signed measure, we must have \( \mu(\emptyset) = 0 \) since

\[
\mu(A) = \mu(A + \emptyset) = \mu(A) + \mu(\emptyset).
\]
We will be measuring these sets using a signed measure (meaning it might be positive or negative). In particular, a real-valued function \( \mu \) defined on \( \mathcal{F}_n \) is called a signed measure if it is set-additive, i.e., for disjoint \( A \) and \( B \), we have

\[
\mu(A \cup B) = \mu(A) + \mu(B)
\]  

(31.2)

For a signed measure, we must have \( \mu(\emptyset) = 0 \) since

\[
\mu(A) = \mu(A + \emptyset) = \mu(A) + \mu(\emptyset).
\]

Note: For sets \( A, B \), set-difference is \( A \setminus B \equiv A \cap B^c \).
We will be measuring these sets using a signed measure (meaning it might be positive or negative). In particular, a real-valued function $\mu$ defined on $\mathcal{F}_n$ is called a signed measure if it is set-additive, i.e., for disjoint $A$ and $B$, we have

$$\mu(A \cup B) = \mu(A) + \mu(B)$$  \hspace{1cm} (31.2)

for a signed measure, we must have $\mu(\emptyset) = 0$ since $\mu(A) = \mu(A + \emptyset) = \mu(A) + \mu(\emptyset)$.

Note: For sets $A, B$, set-difference is $A \setminus B \equiv A \cap B^c$.

Thanks to additively, any signed measure on $\mathcal{F}_n$ is defined by its value on the atoms. I.e., any $\tilde{X} \in \mathcal{F}_n$ can be represented as $\tilde{X} = \bigcup_i Y_i$ where $Y_i$ are appropriately chosen atoms.

$$\mu(\tilde{X}) = \sum_i \mu(Y_i)$$
Example: Consider two sets $\tilde{X}_1, \tilde{X}_2$.
Definitions: signed measure

Example: Consider two sets $\tilde{X}_1, \tilde{X}_2$

Signed measure $\mu$ on $\mathcal{F}_n$ determined by four values on four atoms:

$$\mu(\tilde{X}_1 \cap \tilde{X}_2), \mu(\tilde{X}_1^c \cap \tilde{X}_2), \mu(\tilde{X}_1 \cap \tilde{X}_2^c), \mu(\tilde{X}_1^c \cap \tilde{X}_2^c),$$

(31.3)
Definitions: signed measure

Example: Consider two sets $\tilde{X}_1, \tilde{X}_2$

Signed measure $\mu$ on $\mathcal{F}_n$ determined by four values on four atoms:

$$\mu(\tilde{X}_1 \cap \tilde{X}_2), \mu(\tilde{X}_1^c \cap \tilde{X}_2), \mu(\tilde{X}_1 \cap \tilde{X}_2^c), \mu(\tilde{X}_1^c \cap \tilde{X}_2^c),$$  

(31.3)

For example, we have

$$\mu(\tilde{X}_1) = \mu\left((\tilde{X}_1 \cap \tilde{X}_2^c) \cup (\tilde{X}_1 \cap \tilde{X}_2)\right) = \mu(\tilde{X}_1 \cap \tilde{X}_2^c) + \mu(\tilde{X}_1 \cap \tilde{X}_2)$$  

(31.4)
We said earlier that $\tilde{X}_i$ was associated with a random variable $X_i$. Loosely speaking, the set $\tilde{X}_i$ represents the "uncertainty" or "information" contained within $X_i$. Ex: Two random variables $X_1, X_2$, define the universal set $\Omega = \tilde{X}_1 \cup \tilde{X}_2$ which is the set of everything (for the current $n=2$) $\Rightarrow X_c \equiv \Omega \cap X$ for any $X \in F_n$. By definition of $\Omega$, one of the atoms is always empty, namely $\tilde{X}_c_1 \cap \tilde{X}_c_2 = (\tilde{X}_1 \cup \tilde{X}_2)^c = \emptyset$, so this atom has no area and is not shown in the Venn diagram previously seen. For these two random variables $X_1$ and $X_2$, we already have the Shannon information measures: $H(X_1), H(X_2), H(X_1 | X_2), H(X_2 | X_1), H(X_1, X_2)$. I(\tilde{X}_1; X_2). Let's next associate these with $\mu$ defined on the four atoms.
Random variables

- We said earlier that $\tilde{X}_i$ was associated with a random variable $X_i$.
- Loosely speaking, the set $\tilde{X}_i$ represents the “uncertainty” or “information” contained within $X_i$. 

Examples:

- Two random variables $X_1, X_2$ define the universal set $\Omega = \tilde{X}_1 \cup \tilde{X}_2$, which is the set of everything (for the current $n = 2$).

$X_c \equiv \Omega \cap X$ for any $X \in F_n$.

By definition of $\Omega$, one of the atoms is always empty, namely $\tilde{X}_c_1 \cap \tilde{X}_c_2 = (\tilde{X}_1 \cup \tilde{X}_2)^c = \emptyset$, so this atom has no area and is not shown in the Venn diagram previously seen.

For these two random variables $X_1$ and $X_2$, we already have the Shannon information measures:

$H(X_1)$, $H(X_2)$, $H(X_1 | X_2)$, $H(X_2 | X_1)$, $H(X_1, X_2)$ $I(X_1; X_2)$.

Let's next associate these with $\mu$ defined on the four atoms.
Random variables

- We said earlier that $\tilde{X}_i$ was associated with a random variable $X_i$.
- Loosely speaking, the set $\tilde{X}_i$ represents the “uncertainty” or “information” contained within $X_i$.
- Ex: Two random variables $X_1, X_2$, define the universal set $\Omega = \tilde{X}_1 \cup \tilde{X}_2$ which is the set of everything (for the current $n = 2$) $\Rightarrow X^c \equiv \Omega \setminus X$ for any $X \in \mathcal{F}_n$. 
Random variables

- We said earlier that \( \tilde{X}_i \) was associated with a random variable \( X_i \).
- Loosely speaking, the set \( \tilde{X}_i \) represents the “uncertainty” or “information” contained within \( X_i \).
- Ex: Two random variables \( X_1, X_2 \), define the universal set \( \Omega = \tilde{X}_1 \cup \tilde{X}_2 \) which is the set of everything (for the current \( n = 2 \)) \( \Rightarrow X^c \equiv \Omega \setminus X \) for any \( X \in \mathcal{F}_n \).
- By definition of \( \Omega \), one of the atoms is always empty, namely \( \tilde{X}_1^c \cap \tilde{X}_2^c = (\tilde{X}_1 \cup \tilde{X}_2)^c = \emptyset \), so this atom has no area and is not shown in the Venn diagram previously seen.
Random variables

- We said earlier that $\tilde{X}_i$ was associated with a random variable $X_i$.
- Loosely speaking, the set $\tilde{X}_i$ represents the “uncertainty” or “information” contained within $X_i$.
- Ex: Two random variables $X_1, X_2$, define the universal set $\Omega = \tilde{X}_1 \cup \tilde{X}_2$ which is the set of everything (for the current $n = 2$) $\Rightarrow X^c = \Omega \setminus X$ for any $X \in \mathcal{F}_n$.
- By definition of $\Omega$, one of the atoms is always empty, namely $\tilde{X}_1^c \cap \tilde{X}_2^c = (\tilde{X}_1 \cup \tilde{X}_2)^c = \emptyset$, so this atom has no area and is not shown in the Venn diagram previously seen.
- For these two random variables $X_1$ and $X_2$, we already have the Shannon information measures: $H(X_1), H(X_2), H(X_1|X_2), H(X_2|X_1), H(X_1, X_2) I(X_1; X_2)$. 
We said earlier that \( \tilde{X}_i \) was associated with a random variable \( X_i \).

Loosely speaking, the set \( \tilde{X}_i \) represents the “uncertainty” or “information” contained within \( X_i \).

Ex: Two random variables \( X_1, X_2 \), define the universal set \( \Omega = \tilde{X}_1 \cup \tilde{X}_2 \) which is the set of everything (for the current \( n = 2 \)) \( \Rightarrow X^c \equiv \Omega \setminus X \) for any \( X \in F_n \).

By definition of \( \Omega \), one of the atoms is always empty, namely \( \tilde{X}_1^c \cap \tilde{X}_2^c = (\tilde{X}_1 \cup \tilde{X}_2)^c = \emptyset \), so this atom has no area and is not shown in the Venn diagram previously seen.

For these two random variables \( X_1 \) and \( X_2 \), we already have the Shannon information measures: \( H(X_1), H(X_2), H(X_1|X_2), H(X_2|X_1), H(X_1, X_2) \) \( I(X_1; X_2) \).

Lets next associate these with \( \mu \) defined on the four atoms.
We can make the following defining associations with signed measure $\mu^*$:

1. $\mu^*(\tilde{X}_1 \cap \tilde{X}_2) = I(X_1;X_2)$ \hfill (31.5)
2. $\mu^*(\tilde{X}_1 \cap \tilde{X}_2^c) = \mu^*(\tilde{X}_1 \setminus \tilde{X}_2) = H(X_1|X_2)$ \hfill (31.6)
3. $\mu^*(\tilde{X}_1^c \cap \tilde{X}_2) = \mu^*(\tilde{X}_2 \setminus \tilde{X}_1) = H(X_2|X_1)$ \hfill (31.7)
4. $\mu^*(\tilde{X}_1^c \cap \tilde{X}_2^c) = \mu^*(\emptyset) = 0$ \hfill (31.8)
Signed Measures and Shannon Measures

- We can make the following defining associations with signed measure $\mu^*$:
  
  \begin{align*}
  \mu^*(\tilde{X}_1 \cap \tilde{X}_2) &= I(X_1; X_2) & (31.5) \\
  \mu^*(\tilde{X}_1 \cap \tilde{X}_2^c) &= \mu^*(\tilde{X}_1 \setminus \tilde{X}_2) = H(X_1 | X_2) & (31.6) \\
  \mu^*(\tilde{X}_1^c \cap \tilde{X}_2) &= \mu^*(\tilde{X}_2 \setminus \tilde{X}_1) = H(X_2 | X_1) & (31.7) \\
  \mu^*(\tilde{X}_1^c \cap \tilde{X}_2^c) &= \mu^*(\emptyset) = 0 & (31.8)
  \end{align*}

- We have instantiated the measures of the four atoms with values (could be arbitrary values, but we chose to use entropic quantities).
Signed Measures and Shannon Measures

- We can make the following defining associations with signed measure $\mu^*$:

$$\mu^*(\tilde{X}_1 \cap \tilde{X}_2) = I(X_1; X_2)$$  \hspace{1cm} (31.5) \\
$$\mu^*(\tilde{X}_1 \cap \tilde{X}_2^c) = \mu^*(\tilde{X}_1 \setminus \tilde{X}_2) = H(X_1|X_2)$$  \hspace{1cm} (31.6) \\
$$\mu^*(\tilde{X}_1^c \cap \tilde{X}_2) = \mu^*(\tilde{X}_2 \setminus \tilde{X}_1) = H(X_2|X_1)$$  \hspace{1cm} (31.7) \\
$$\mu^*(\tilde{X}_1^c \cap \tilde{X}_2^c) = \mu^*(\emptyset) = 0$$  \hspace{1cm} (31.8)

- We have instantiated the measures of the four atoms with values (could be arbitrary values, but we chose to use entropic quantities).

- Given these definitions, what would $\mu^*(\tilde{X}_1), \mu^*(\tilde{X}_2)$, and $\mu^*(\tilde{X}_1 \cup \tilde{X}_2)$ be?
Signed Measures and Shannon Measures

- We can make the following defining associations with signed measure $\mu^*$:

$$\mu^*(\tilde{X}_1 \cap \tilde{X}_2) = I(X_1; X_2) \quad (31.5)$$

$$\mu^*(\tilde{X}_1 \cap \tilde{X}_2^c) = \mu^*(\tilde{X}_1 \setminus \tilde{X}_2) = H(X_1|X_2) \quad (31.6)$$

$$\mu^*(\tilde{X}_1^c \cap \tilde{X}_2) = \mu^*(\tilde{X}_2 \setminus \tilde{X}_1) = H(X_2|X_1) \quad (31.7)$$

$$\mu^*(\tilde{X}_1^c \cap \tilde{X}_2^c) = \mu^*(\emptyset) = 0 \quad (31.8)$$

- We have instantiated the measures of the four atoms with values (could be arbitrary values, but we chose to use entropic quantities).

- Given these definitions, what would $\mu^*(\tilde{X}_1)$, $\mu^*(\tilde{X}_2)$, and $\mu^*(\tilde{X}_1 \cup \tilde{X}_2)$ be?

$$\mu^*(\tilde{X}_1)$$
We can make the following defining associations with signed measure $\mu^*$:

\[ \mu^*(\tilde{X}_1 \cap \tilde{X}_2) = I(X_1; X_2) \]  \hspace{1cm} (31.5)

\[ \mu^*(\tilde{X}_1 \cap \tilde{X}_2^c) = \mu^*(\tilde{X}_1 \setminus \tilde{X}_2) = H(X_1|X_2) \]  \hspace{1cm} (31.6)

\[ \mu^*(\tilde{X}_1^c \cap \tilde{X}_2) = \mu^*(\tilde{X}_2 \setminus \tilde{X}_1) = H(X_2|X_1) \]  \hspace{1cm} (31.7)

\[ \mu^*(\tilde{X}_1^c \cap \tilde{X}_2^c) = \mu^*(\emptyset) = 0 \]  \hspace{1cm} (31.8)

We have instantiated the measures of the four atoms with values (could be arbitrary values, but we chose to use entropic quantities).

Given these definitions, what would $\mu^*(\tilde{X}_1)$, $\mu^*(\tilde{X}_2)$, and $\mu^*(\tilde{X}_1 \cup \tilde{X}_2)$ be?

\[ \mu^*(\tilde{X}_1) = \mu^*((\tilde{X}_1 \cap \tilde{X}_2) \cup (\tilde{X}_1 \cap \tilde{X}_2^c)) \]  \hspace{1cm} (31.9)
Signed Measures and Shannon Measures

- We can make the following defining associations with signed measure $\mu^*$:

\[
\begin{align*}
\mu^*(\tilde{X}_1 \cap \tilde{X}_2) &= I(X_1; X_2) \quad (31.5) \\
\mu^*(\tilde{X}_1 \cap \tilde{X}_2^c) &= \mu^*(\tilde{X}_1 \setminus \tilde{X}_2) = H(X_1|X_2) \quad (31.6) \\
\mu^*(\tilde{X}_1^c \cap \tilde{X}_2) &= \mu^*(\tilde{X}_2 \setminus \tilde{X}_1) = H(X_2|X_1) \quad (31.7) \\
\mu^*(\tilde{X}_1^c \cap \tilde{X}_2^c) &= \mu^*(\emptyset) = 0 \quad (31.8)
\end{align*}
\]

- We have instantiated the measures of the four atoms with values (could be arbitrary values, but we chose to use entropic quantities).

- Given these definitions, what would $\mu^*(\tilde{X}_1)$, $\mu^*(\tilde{X}_2)$, and $\mu^*(\tilde{X}_1 \cup \tilde{X}_2)$ be?

\[
\begin{align*}
\mu^*(\tilde{X}_1) &= \mu^*((\tilde{X}_1 \cap \tilde{X}_2) \cup (\tilde{X}_1 \cap \tilde{X}_2^c)) = I(X_1; X_2) + H(X_1|X_2) = H(X_1) \quad (31.9, 31.10)
\end{align*}
\]
Signed Measures and Shannon Measures

What would \( \mu^*(\tilde{X}_2) \) be?

(31.11)
What would $\mu^*(\tilde{X}_2)$ be?

$$\mu^*(\tilde{X}_2) = H(X_2)$$  \hspace{1cm} (31.11)
Signed Measures and Shannon Measures

What would $\mu^*(\tilde{X}_2)$ be?

$$\mu^*(\tilde{X}_2) = H(X_2)$$  \hfill (31.11)

What about $\mu^*(\tilde{X}_1 \cup \tilde{X}_2)$?

$$\mu^*(\tilde{X}_1 \cup \tilde{X}_2) = \mu((\bigcup \text{all atoms} A) \cup \tilde{X}_2)$$  \hfill (31.12)

$$= I(X_1;X_2) + H(X_1|X_2) + H(X_2|X_1) + 0$$  \hfill (31.13)

$$= H(X_1, X_2)$$  \hfill (31.14)
What would $\mu^*(\tilde{X}_2)$ be?

$$\mu^*(\tilde{X}_2) = H(X_2) \quad (31.11)$$

What about $\mu^*(\tilde{X}_1 \cup \tilde{X}_2)$?

$$\mu^*(\tilde{X}_1 \cup \tilde{X}_2) = H(X_1, X_2) \quad (31.14)$$
Signed Measures and Shannon Measures

- What would \( \mu^*(\tilde{X}_2) \) be?
  \[
  \mu^*(\tilde{X}_2) = H(X_2)
  \]  
  (31.11)

- What about \( \mu^*(\tilde{X}_1 \cup \tilde{X}_2) \)?
  \[
  \mu^*(\tilde{X}_1 \cup \tilde{X}_2) = \mu\left( \bigcup_{\text{all atoms } A} A \right)
  \]  
  (31.12)
  
  \[
  = H(X_1, X_2)
  \]  
  (31.14)
Signed Measures and Shannon Measures

- What would $\mu^*(\tilde{X}_2)$ be?
  \[
  \mu^*(\tilde{X}_2) = H(X_2)
  \]  
  (31.11)

- What about $\mu^*(\tilde{X}_1 \cup \tilde{X}_2)$?
  \[
  \mu^*(\tilde{X}_1 \cup \tilde{X}_2) = \mu \left( \bigcup_{\text{all atoms } A} A \right)
  = I(X_1; X_2) + H(X_1|X_2) + H(X_2|X_1) + 0
  \]  
  (31.12)  
  (31.13)  
  (31.14)
Signed Measures and Shannon Measures

- What would $\mu^*(\tilde{X}_2)$ be?

$$\mu^*(\tilde{X}_2) = H(X_2) \quad (31.11)$$

- What about $\mu^*(\tilde{X}_1 \cup \tilde{X}_2)$?

$$\mu^*(\tilde{X}_1 \cup \tilde{X}_2) = \mu\left( \bigcup_{\text{all atoms } A} A \right) = I(X_1; X_2) + H(X_1 | X_2) + H(X_2 | X_1) + 0 \quad (31.13)$$

$$= H(X_1, X_2) \quad (31.14)$$
Signed Measures and Shannon Measures

- What would $\mu^*(\tilde{X}_2)$ be?

  $$\mu^*(\tilde{X}_2) = H(X_2) \quad (31.11)$$

- What about $\mu^*(\tilde{X}_1 \cup \tilde{X}_2)$?

  $$\mu^*(\tilde{X}_1 \cup \tilde{X}_2) = \mu \left( \bigcup_{\text{all atoms } A} A \right) \quad (31.12)$$

  $$= I(X_1; X_2) + H(X_1 | X_2) + H(X_2 | X_1) + 0 \quad (31.13)$$

  $$= H(X_1, X_2) \quad (31.14)$$

- So, we have defined $\mu^*$ only on the atoms, and from this we have, using the signed measure property and set theory, fully recovered all the rest of the Shannon information values.
Unions of sets

What if we define $\mu^*$ only on the unions of sets. I.e., we make the following defining associations:

$$
\mu^*(\emptyset) = 0 \quad (31.15)
$$

$$
\mu^*(\tilde{X}_1) = H(X_1) \quad (31.16)
$$

$$
\mu^*(\tilde{X}_2) = H(X_2) \quad (31.17)
$$

$$
\mu^*(\tilde{X}_1 \cup \tilde{X}_2) = H(X_1, X_2) \quad (31.18)
$$
What if we define $\mu^*$ only on the unions of sets. I.e., we make the following defining associations:

$$\mu^*(\emptyset) = 0$$  \hspace{1cm} (31.15)

$$\mu^*(\tilde{X}_1) = H(X_1)$$  \hspace{1cm} (31.16)

$$\mu^*(\tilde{X}_2) = H(X_2)$$  \hspace{1cm} (31.17)

$$\mu^*(\tilde{X}_1 \cup \tilde{X}_2) = H(X_1, X_2)$$  \hspace{1cm} (31.18)

Then from this, we can (using set theory) get the rest of the values, $I(X_1; X_2), H(X_1|X_2), H(X_2|X_1)$. 
What if we define $\mu^*$ only on the unions of sets. I.e., we make the following defining associations:

\begin{align*}
\mu^*(\emptyset) &= 0 \\
\mu^*(\tilde{X}_1) &= H(X_1) \\
\mu^*(\tilde{X}_2) &= H(X_2) \\
\mu^*(\tilde{X}_1 \cup \tilde{X}_2) &= H(X_1, X_2)
\end{align*}

Then from this, we can (using set theory) get the rest of the values, $I(X_1; X_2)$, $H(X_1|X_2)$, $H(X_2|X_1)$.

E.g., we get:

\begin{align*}
\mu(\tilde{X}_1 \cap \tilde{X}_2) &= \mu(\tilde{X}_1) + \mu(\tilde{X}_2) - \mu(\tilde{X}_1 \cup \tilde{X}_2) \\
&= H(X_1) + H(X_2) - H(X_1, X_2) = I(X_1; X_2)
\end{align*}
So we have recovered Shannon's information measures with the following correspondence:

\[ H/I \leftrightarrow \mu^* \]  
\[ , \leftrightarrow \cup \]  
\[ ; \leftrightarrow \cap \]  
\[ \mid \leftrightarrow \setminus \text{ / set minus} \]  

Note: with measure notation, we could identify \( H(X; Y) = I(X; Y) \), where \( H(X; Y) \) vs. \( H(X, Y) \) would be distinguished only by a semicolon rather than a comma.
So we have recovered Shannon’s information measures with the following correspondence:

\[ H/I \leftrightarrow \mu^* \quad (31.21) \]

, \leftrightarrow \cup \quad (31.22)

; \leftrightarrow \cap \quad (31.23)

| \leftrightarrow \setminus \quad //set minus \quad (31.24)

Note: with measure notation, no distinction between \( H \) and \( I \), we could identify \( H(X;Y) = I(X;Y) \), where \( H(X;Y) \) vs. \( H(X,Y) \) would be distinguished only by a semicolon rather than a comma.
Recovering Shannon

- So we have recovered Shannon’s information measures with the following correspondence:

\[ H/I \longleftrightarrow \mu^* \quad (31.21) \]

\[ , \longleftrightarrow \cup \quad (31.22) \]

\[ ; \longleftrightarrow \cap \quad (31.23) \]

\[ | \longleftrightarrow \setminus \quad //set \ minus \ (31.24) \]

- Note: with measure notation, no distinction between \( H \) and \( I \), we could identify \( H(X;Y) = I(X;Y) \), where \( H(X;Y) \) vs. \( H(X,Y) \) would be distinguished only by a semicolon rather than a comma. **but we don’t do that.**
Recovering Shannon

So we have recovered Shannon’s information measures with the following correspondence:

\[ H/I \longleftrightarrow \mu^* \quad (31.21) \]
\[ , \longleftrightarrow \cup \quad (31.22) \]
\[ ; \longleftrightarrow \cap \quad (31.23) \]
\[ | \longleftrightarrow \setminus \quad \text{//set minus} \quad (31.24) \]

Note: with measure notation, no distinction between \( H \) and \( I \), we could identify \( H(X;Y) = I(X;Y) \), where \( H(X;Y) \) vs. \( H(X,Y) \) would be distinguished only by a semicolon rather than a comma. But we don’t do that.

Does this generalize? The hope is that if there is an information theoretic identity, it would occur iff there is a set theory identity.
For example, a simple example of a well-known property in set theory: The inclusion-exclusion formula for measures, is as follows:

$$\mu(A \cup B) = \mu(A) + \mu(B) - \mu(A \cap B)$$  \hfill (31.25)

which follows since

$$\mu(A \cup B) = \mu(A) + \mu(B \cap A)$$ \hfill (31.28)
For example, a simple example of a well-known property in set theory: The inclusion-exclusion formula for measures, is as follows:

\[ \mu(A \cup B) = \mu(A) + \mu(B) - \mu(A \cap B) \]  

which follows since

\[ \mu(A \cup B) \]  

(31.28)
Recovering Shannon

For example, a simple example of a well-known property in set theory: The inclusion-exclusion formula for measures, is as follows:

\[ \mu(A \cup B) = \mu(A) + \mu(B) - \mu(A \cap B) \]  

(31.25)

which follows since

\[ \mu(A \cup B) = \mu(A \cup (B \setminus A)) \]  

(31.28)
Recovering Shannon

For example, a simple example of a well-known property in set theory: The inclusion-exclusion formula for measures, is as follows:

\[ \mu(A \cup B) = \mu(A) + \mu(B) - \mu(A \cap B) \]  \hspace{1cm} (31.25)

which follows since

\[ \mu(A \cup B) = \mu(A \cup (B \setminus A)) = \mu(A) + \mu(B \setminus A) \]  \hspace{1cm} (31.26)
For example, a simple example of a well-known property in set theory: The inclusion-exclusion formula for measures, is as follows:

\[ \mu(A \cup B) = \mu(A) + \mu(B) - \mu(A \cap B) \]  

(31.25)

which follows since

\[ \mu(A \cup B) = \mu(A \cup (B \setminus A)) = \mu(A) + \mu(B \setminus A) \]  

(31.26)

\[ = \mu(A) + \mu(B \setminus A) + \mu(A \cap B) - \mu(A \cap B) \]  

(31.27)

(31.28)
For example, a simple example of a well-known property in set theory: The inclusion-exclusion formula for measures, is as follows:

\[ \mu(A \cup B) = \mu(A) + \mu(B) - \mu(A \cap B) \]  

which follows since

\[ \mu(A \cup B) = \mu(A \cup (B \setminus A)) = \mu(A) + \mu(B \setminus A) \]  
\[ = \mu(A) + \mu(B \setminus A) + \mu(A \cap B) - \mu(A \cap B) \]  
\[ = \mu(A) + \mu(B) - \mu(A \cap B) \]
For example, a simple example of a well-known property in set theory: The inclusion-exclusion formula for measures, is as follows:

\[ \mu(A \cup B) = \mu(A) + \mu(B) - \mu(A \cap B) \] (31.25)

which follows since

\[ \mu(A \cup B) = \mu(A \cup (B \setminus A)) = \mu(A) + \mu(B \setminus A) \] (31.26)
\[ = \mu(A) + \mu(A \cap B) - \mu(A \cap B) \] (31.27)
\[ = \mu(A) + \mu(B) - \mu(A \cap B) \] (31.28)

Equating our measures, we see that the entropy/mutual information formula is just inclusion-exclusion:

\[ H(X_1, X_2) = H(X_1) + H(X_2) - I(X_1; X_2) \] (31.29)
General case, \( n \geq 2 \)

- \( n \) random variables \( X_1, \ldots, X_n \) corresponding to sets \( \tilde{X}_i \), and with \([n] = \{1, 2, \ldots, n\}\) the index set.
General case, $n \geq 2$

- $n$ random variables $X_1, \ldots, X_n$ corresponding to sets $\tilde{X}_i$, and with $[n] = \{1, 2, \ldots, n\}$ the index set.
- $\Omega = \bigcup_i \tilde{X}_i$ is the universe, and empty atom again is:

\[
A_0 = \bigcap_{i \in [n]} \tilde{X}_i^c = \left( \bigcup_i \tilde{X}_i \right)^c = \emptyset
\]  

(31.30)
General case, \( n \geq 2 \)

- \( n \) random variables \( X_1, \ldots, X_n \) corresponding to sets \( \tilde{X}_i \), and with \( [n] = \{1, 2, \ldots, n\} \) the index set.
- \( \Omega = \bigcup_i \tilde{X}_i \) is the universe, and empty atom again is:
\[
A_0 = \bigcap_{i \in [n]} \tilde{X}_i^c = \left( \bigcup_i \tilde{X}_i \right)^c = \emptyset
\]  
(31.30)

- Non-empty atoms are \( \mathcal{A} \triangleq \{\text{all atoms}\} \setminus \{A_0\} \) which are those that are not assuredly empty, so \( |\mathcal{A}| = 2^n - 1 \).
General case, \( n \geq 2 \)

- \( n \) random variables \( X_1, \ldots, X_n \) corresponding to sets \( \tilde{X}_i \), and with 
  \( [n] = \{1, 2, \ldots, n\} \) the index set.
- \( \Omega = \bigcup_i \tilde{X}_i \) is the universe, and empty atom again is:

\[
A_0 = \bigcap_{i \in [n]} \tilde{X}_i^c = \left( \bigcup_i \tilde{X}_i \right)^c = \emptyset \tag{31.30}
\]

- Non-empty atoms are \( \mathcal{A} \triangleq \{ \text{all atoms} \} \setminus \{ A_0 \} \) which are those that are not assuredly empty, so \( |\mathcal{A}| = 2^n - 1 \).
- When values of \( \mu(\cdot) \) is given for all \( \mathcal{A} \), then this defines \( \mu(\cdot) \) on all \( \mathcal{F}_n \). Why?
**General case, \( n \geq 2 \)**

- \( n \) random variables \( X_1, \ldots, X_n \) corresponding to sets \( \tilde{X}_i \), and with \( [n] = \{1, 2, \ldots, n\} \) the index set.
- \( \Omega = \bigcup_i \tilde{X}_i \) is the universe, and empty atom again is:
  \[
  A_0 = \bigcap_{i \in [n]} \tilde{X}_i^c = \left( \bigcup_i \tilde{X}_i \right)^c = \emptyset \quad (31.30)
  \]
- Non-empty atoms are \( \mathcal{A} \triangleq \{ \text{all atoms} \} \setminus \{ A_0 \} \) which are those that are not assuredly empty, so \( |\mathcal{A}| = 2^n - 1 \).
- When values of \( \mu(\cdot) \) is given for all \( \mathcal{A} \), then this defines \( \mu(\cdot) \) on all \( \mathcal{F}_n \). Why? Since the atoms \( \mathcal{A} \) partition and hence fully cover \( \Omega \).
General case, $n \geq 2$

- $n$ random variables $X_1, \ldots, X_n$ corresponding to sets $\tilde{X}_i$, and with $[n] = \{1, 2, \ldots, n\}$ the index set.
- $\Omega = \bigcup_i \tilde{X}_i$ is the universe, and empty atom again is:

$$A_0 = \bigcap_{i \in [n]} \tilde{X}_i^c = \left( \bigcup_i \tilde{X}_i \right)^c = \emptyset$$

(31.30)

- Non-empty atoms are $\mathcal{A} \triangleq \{\text{all atoms}\} \setminus \{A_0\}$ which are those that are not assuredly empty, so $|\mathcal{A}| = 2^n - 1$.
- When values of $\mu(\cdot)$ is given for all $\mathcal{A}$, then this defines $\mu(\cdot)$ on all $\mathcal{F}_n$. Why? Since the atoms $\mathcal{A}$ partition and hence fully cover $\Omega$.
- Notation: for $G \subseteq [n]$, $X_G = (X_i, i \in G)$ for index set $G$. 


**General case, \( n \geq 2 \)**

- \( n \) random variables \( X_1, \ldots, X_n \) corresponding to sets \( \tilde{X}_i \), and with \( [n] = \{1, 2, \ldots, n\} \) the index set.
- \( \Omega = \bigcup_i \tilde{X}_i \) is the universe, and empty atom again is:

\[
A_0 = \bigcap_{i \in [n]} \tilde{X}_i^c = \left( \bigcup_i \tilde{X}_i \right)^c = \emptyset \quad (31.30)
\]

- Non-empty atoms are \( \mathcal{A} \triangleq \{ \text{all atoms} \} \setminus \{ A_0 \} \) which are those that are not assuredly empty, so \( |\mathcal{A}| = 2^n - 1 \).
- When values of \( \mu(\cdot) \) is given for all \( \mathcal{A} \), then this defines \( \mu(\cdot) \) on all \( \mathcal{F}_n \). Why? Since the atoms \( \mathcal{A} \) partition and hence fully cover \( \Omega \).
- Notation: for \( G \subseteq [n] \), \( X_G = (X_i, i \in G) \) for index set \( G \).
- Notation: for \( G \subseteq [n] \), \( \tilde{X}_G = \bigcup_{i \in G} \tilde{X}_i \).
General case, $n \geq 2$

- $n$ random variables $X_1, \ldots, X_n$ corresponding to sets $\tilde{X}_i$, and with $[n] = \{1, 2, \ldots, n\}$ the index set.
- $\Omega = \bigcup_i \tilde{X}_i$ is the universe, and empty atom again is:

$$A_0 = \bigcap_{i \in [n]} \tilde{X}_i^c = \left(\bigcup_i \tilde{X}_i\right)^c = \emptyset \quad (31.30)$$

- Non-empty atoms are $\mathcal{A} \triangleq \{\text{all atoms}\} \setminus \{A_0\}$ which are those that are not assuredly empty, so $|\mathcal{A}| = 2^n - 1$.
- When values of $\mu(\cdot)$ is given for all $\mathcal{A}$, then this defines $\mu(\cdot)$ on all $\mathcal{F}_n$. Why? Since the atoms $\mathcal{A}$ partition and hence fully cover $\Omega$.
- Notation: for $G \subseteq [n]$, $X_G = (X_i, i \in G)$ for index set $G$.
- Notation: for $G \subseteq [n]$, $\tilde{X}_G = \bigcup_{i \in G} \tilde{X}_i$.
- Definition, non-empty unions (note strictness on left side):

$$\mathcal{B} \triangleq \left\{\tilde{X}_G : \emptyset \subset G \subseteq [n]\right\} \quad (31.31)$$
\[ x_1, x_2, \ldots \]
Theorem 31.3.1

Signed measure $\mu$ on $\mathcal{F}_n$ is fully specified by $\{\mu(B) : B \in \mathcal{B}\}$ which can be any set of real numbers.

- So before, we defined $\mu$ on all atoms and noted that this defined $\mu$ everywhere else.
What needs to be specified

Theorem 31.3.1

Signed measure $\mu$ on $\mathcal{F}_n$ is fully specified by $\{\mu(B) : B \in \mathcal{B}\}$ which can be any set of real numbers.

- So before, we defined $\mu$ on all atoms and noted that this defined $\mu$ everywhere else.
- Here, in the above theorem, we are defining $\mu$ only on elements of $\mathcal{B}$ and are again saying that this defines the values of $\mu$ everywhere else.
What needs to be specified

**Theorem 31.3.1**

Signed measure $\mu$ on $\mathcal{F}_n$ is fully specified by $\{\mu(B) : B \in \mathcal{B}\}$ which can be any set of real numbers.

- So before, we defined $\mu$ on all atoms and noted that this defined $\mu$ everywhere else.

- Here, in the above theorem, we are defining $\mu$ only on elements of $\mathcal{B}$ and are again saying that this defines the values of $\mu$ everywhere else.

- We will see that this allows us to generate all standard mutual-information quantities, and some other less standard ones, using just entropy to fill out $\mu$ on $\mathcal{B}$. 
Consider a simple signed measure $\mu(A) = |A|$, the cardinality (or size or counting) measure (but the same idea works for any signed measure)
Consider a simple signed measure \( \mu(A) = |A| \), the cardinality (or size or counting) measure (but the same idea works for any signed measure).

First note, from the binomial expansion:

\[
0 = 1 - 1 = (1 - 1)^n = \sum_{\ell=0}^{n} \binom{n}{\ell} (-1)^\ell
\]
Consider a simple signed measure $\mu(A) = |A|$, the cardinality (or size or counting) measure (but the same idea works for any signed measure)

First note, from the binomial expansion:

$$0 = 1 - 1$$
Consider a simple signed measure $\mu(A) = |A|$, the cardinality (or size or counting) measure (but the same idea works for any signed measure).

First note, from the binomial expansion:

\[ 0 = 1 - 1 = (1 - 1)^n \]  

(31.32)
Consider a simple signed measure \( \mu(A) = |A| \), the cardinality (or size or counting) measure (but the same idea works for any signed measure)

First note, from the binomial expansion:

\[
0 = 1 - 1 = (1 - 1)^n = \sum_{\ell=0}^{n} \binom{n}{\ell} (-1)^\ell (1)^{n-\ell}
\]

(31.32) (31.33)
Consider a simple signed measure $\mu(A) = |A|$, the cardinality (or size or counting) measure (but the same idea works for any signed measure).

First note, from the binomial expansion:

$$0 = 1 - 1 = (1 - 1)^n$$

$$= \sum_{\ell=0}^{n} \binom{n}{\ell} (-1)^{\ell} (1)^{n-\ell}$$

$$= 1 - \binom{n}{1} + \binom{n}{2} - \cdots + (-1)^n \binom{n}{n}$$
Consider a simple signed measure $\mu(A) = |A|$, the cardinality (or size or counting) measure (but the same idea works for any signed measure).

First note, from the binomial expansion:

$$0 = 1 - 1 = (1 - 1)^n = \sum_{\ell=0}^{n} \binom{n}{\ell} (-1)^\ell (1)^{n-\ell}$$

$$= 1 - \binom{n}{1} + \binom{n}{2} - \ldots + (-1)^n \binom{n}{n}$$

We have $n$ sets $A_i$, for $i = 1 \ldots n$ such that $A_i \subseteq \Omega$. 
We have \( n \) sets \( A_i \), for \( i = 1 \ldots n \) such that \( A_i \subseteq \Omega \). What we wish to prove is the form of the inclusion/exclusion formula.

\[
| \cap_{i=1}^n A_i | = \sum_{i=1}^n |A_i| - \sum_{1 \leq i < j \leq n} |A_i \cup A_j| + \sum_{1 \leq i < j < k \leq n} |A_i \cup A_j \cup A_k| + \ldots
\]

[31.35]

\[
+ (-1)^{n-1} |A_1 \cup A_2 \cup \ldots \cup A_n|
\]

[31.37]
Inclusion/Exclusion

- We have \( n \) sets \( A_i \), for \( i = 1 \ldots n \) such that \( A_i \subseteq \Omega \). What we wish to prove is the form of the inclusion/exclusion formula.

\[
| \bigcap_{i=1}^{n} A_i | = \sum_{i=1}^{n} |A_i| - \sum_{1 \leq i < j \leq n} |A_i \cup A_j| + \sum_{1 \leq i < j < k \leq n} |A_i \cup A_j \cup A_k| + \ldots + (-1)^{n-1} |A_1 \cup A_2 \cup \ldots \cup A_n|
\]  

(31.35)

(31.36)

(31.37)

- Note the pattern: first we over count, then we undercount, and then overcount, etc. ... until the last term finally fixes things.
We have \( n \) sets \( A_i \), for \( i = 1 \ldots n \) such that \( A_i \subseteq \Omega \). What we wish to prove is the form of the inclusion/exclusion formula.

\[
| \cap_{i=1}^{n} A_i | = \sum_{i=1}^{n} |A_i| - \sum_{1 \leq i < j \leq n} |A_i \cup A_j| + \sum_{1 \leq i < j < k \leq n} |A_i \cup A_j \cup A_k| + \ldots + (-1)^{n-1} |A_1 \cup A_2 \cup \ldots \cup A_n| \tag{31.35}
\]

Note the pattern: first we over count, then we undercount, and then overcount, etc. \ldots until the last term finally fixes things.

Special case of sieve methods: general mathematical methods to count sizes of sets of integers (e.g., sieve of Eratosthenes).
Consider an \( x \in \Omega \) where \( x \in A_i \) for all \( i = 1 \ldots n \).
Consider an \( x \in \Omega \) where \( x \in A_i \) for all \( i = 1 \ldots n \).

Then the l.h.s. of Equation 31.37 contributes only 1 (unity) for this \( x \).
Consider an $x \in \Omega$ where $x \in A_i$ for all $i = 1 \ldots n$.

Then the l.h.s. of Equation 31.37 contributes only 1 (unity) for this $x$.

For the r.h.s. of Equation 31.37, we want to look what the contribution for this particular $x$ will be, and we do this on the next slide.
For the r.h.s. of Equation 31.37, we want to look what the contribution for this particular $x$ will be, and it is:

$$n - \binom{n}{2} + \binom{n}{3} - \binom{n}{4} + \ldots + (-1)^{n-1} \binom{n}{n}$$

(31.38)

Thus, the contribution for any particular $x$ is one both on the l.h.s. and on the r.h.s. of the equation.
For the r.h.s. of Equation 31.37, we want to look what the contribution for this particular $x$ will be, and it is:

$$n - \binom{n}{2} + \binom{n}{3} - \binom{n}{4} + \ldots + (-1)^{n-1} \binom{n}{n}$$  

(31.38)

$$= (-1)^0 \binom{n}{1} + (-1)^1 \binom{n}{2} + (-1)^2 \binom{n}{3} + (-1)^3 \binom{n}{4} + \ldots + (-1)^{n-1} \binom{n}{n}$$  

(31.39)

Thus, the contribution for any particular $x$ is one both on the l.h.s. and on the r.h.s. of the equation.
For the r.h.s. of Equation 31.37, we want to look what the contribution for this particular \( x \) will be, and it is:

\[
n - \binom{n}{2} + \binom{n}{3} - \binom{n}{4} + \ldots + (-1)^{n-1} \binom{n}{n}
\]

\[
= (-1)^0 \binom{n}{1} + (-1)^1 \binom{n}{2} + (-1)^2 \binom{n}{3} + (-1)^3 \binom{n}{4} + \ldots + (-1)^{n-1} \binom{n}{n}
\]

\[
= (-1)^0 \left[ (-1)^1 \binom{n}{1} + (-1)^2 \binom{n}{2} + (-1)^3 \binom{n}{3} + (-1)^4 \binom{n}{4} + \ldots + (-1)^n \binom{n}{n} \right]
\]

Thus, the contribution for any particular \( x \) is one both on the l.h.s. and on the r.h.s. of the equation.
For the r.h.s. of Equation 31.37, we want to look what the contribution for this particular \( x \) will be, and it is:

\[
n - \binom{n}{2} + \binom{n}{3} - \binom{n}{4} + \ldots + (-1)^{n-1} \binom{n}{n}
\]

(31.38)

\[
= (-1)^0 \binom{n}{1} + (-1)^1 \binom{n}{2} + (-1)^2 \binom{n}{3} + (-1)^3 \binom{n}{4} + \ldots + (-1)^{n-1} \binom{n}{n}
\]

(31.39)

\[
= (-1) \left[ (-1)^1 \binom{n}{1} + (-1)^2 \binom{n}{2} + (-1)^3 \binom{n}{3} + (-1)^4 \binom{n}{4} + \ldots + (-1)^n \binom{n}{n} \right]
\]

(31.40)

\[
= (-1) \left[ (-1)^{1} \binom{n}{1} + (-1)^2 \binom{n}{2} + (-1)^3 \binom{n}{3} + (-1)^4 \binom{n}{4} + \ldots + (-1)^n \binom{n}{n} + (1 - 1) \right]
\]

(31.41)

Thus, the contribution for any particular \( x \) is one both on the l.h.s. and on the r.h.s. of the equation.
For the r.h.s. of Equation 31.37, we want to look what the contribution for this particular \( x \) will be, and it is:

\[
n - \binom{n}{2} + \binom{n}{3} - \binom{n}{4} + \ldots + (-1)^{n-1} \binom{n}{n}
\]  

(31.38)

\[
= (-1)^0 \binom{n}{1} + (-1)^1 \binom{n}{2} + (-1)^2 \binom{n}{3} + (-1)^3 \binom{n}{4} + \ldots + (-1)^{n-1} \binom{n}{n}
\]  

(31.39)

\[
= (-1) \left[ (-1)^1 \binom{n}{1} + (-1)^2 \binom{n}{2} + (-1)^3 \binom{n}{3} + (-1)^4 \binom{n}{4} + \ldots + (-1)^n \binom{n}{n} \right]
\]  

(31.40)

\[
= (-1) \left[ (-1)^0 \binom{n}{0} + (-1)^1 \binom{n}{1} + (-1)^2 \binom{n}{2} + (-1)^3 \binom{n}{3} + (-1)^4 \binom{n}{4} + \ldots + (-1)^n \binom{n}{n} - 1 \right]
\]  

(31.41)

Thus, the contribution for any particular \( x \) is one both on the l.h.s. and on the r.h.s. of the equation.
For the r.h.s. of Equation 31.37, we want to look what the contribution for this particular \( x \) will be, and it is:

\[
\begin{align*}
&= n - \binom{n}{2} + \binom{n}{3} - \binom{n}{4} + \ldots + (-1)^{n-1} \binom{n}{n} \\
&= (-1)^0 \binom{n}{1} + (-1)^1 \binom{n}{2} + (-1)^2 \binom{n}{3} + (-1)^3 \binom{n}{4} + \ldots + (-1)^{n-1} \binom{n}{n} \\
&= (-1) \left[ (-1)^1 \binom{n}{1} + (-1)^2 \binom{n}{2} + (-1)^3 \binom{n}{3} + (-1)^4 \binom{n}{4} + \ldots + (-1)^n \binom{n}{n} \right] \\
&= (-1) \left[ (-1)^0 \binom{n}{0} + (-1)^1 \binom{n}{1} + (-1)^2 \binom{n}{2} + (-1)^3 \binom{n}{3} + (-1)^4 \binom{n}{4} + \ldots + (-1)^n \binom{n}{n} - 1 \right] \\
&= (-1) \left[ (1 - 1)^n - 1 \right]
\end{align*}
\]

Thus, the contribution for any particular \( x \) is one both on the l.h.s. and on the r.h.s. of the equation.
For the r.h.s. of Equation 31.37, we want to look what the contribution for this particular \( x \) will be, and it is:

\[
n - \binom{n}{2} + \binom{n}{3} - \binom{n}{4} + \ldots + (-1)^{n-1} \binom{n}{n}
\]

(31.38)

\[
= (-1)^0 \binom{n}{1} + (-1)^1 \binom{n}{2} + (-1)^2 \binom{n}{3} + (-1)^3 \binom{n}{4} + \ldots + (-1)^{n-1} \binom{n}{n}
\]

(31.39)

\[
= (-1) \left[ (-1)^1 \binom{n}{1} + (-1)^2 \binom{n}{2} + (-1)^3 \binom{n}{3} + (-1)^4 \binom{n}{4} + \ldots + (-1)^n \binom{n}{n} \right]
\]

(31.40)

\[
= (-1) \left[ (-1)^1 \binom{n}{1} + (-1)^2 \binom{n}{2} + (-1)^3 \binom{n}{3} + (-1)^4 \binom{n}{4} + \ldots + (-1)^n \binom{n}{n} + (1 - 1) \right]
\]

(31.41)

\[
= (-1) \left[ (-1)^0 \binom{n}{0} + (-1)^1 \binom{n}{1} + (-1)^2 \binom{n}{2} + (-1)^3 \binom{n}{3} + (-1)^4 \binom{n}{4} + \ldots + (-1)^n \binom{n}{n} - 1 \right]
\]

(31.42)

\[
= (-1) \left( (1 - 1)^n - 1 \right) = 1
\]

(31.43)

Thus, the contribution for any particular \( x \) is one both on the l.h.s. and on the r.h.s. of the equation.
Next, suppose that \( x \in A_i \) for \( i \in S \), where \( |S| = k \). In other words, \( x \) is in only exactly \( k \) of the sets \( A_k \) rather than all of them, where \( 0 \leq k < n \).
Next, suppose that $x \in A_i$ for $i \in S$, where $|S| = k$. In other words, $x$ is in only exactly $k$ of the sets $A_k$ rather than all of them, where $0 \leq k < n$.

**Reminder of the inclusion/exclusion formula.**

\[
| \cap_{i=1}^{n} A_i | = \sum_{i=1}^{n} |A_i| - \sum_{1 \leq i < j \leq n} |A_i \cup A_j| + \sum_{1 \leq i < j < k \leq n} |A_i \cup A_j \cup A_k| + \ldots + (-1)^{n-1} |A_1 \cup A_2 \cup \ldots \cup A_n| \tag{31.44}
\]

\[
\text{The l.h.s. of Equation 31.46 now contributes only 0 (zero) for this } x. \text{ For the r.h.s. of Equation 31.46, then the contribution for this particular } x \text{ will be . . .}
\]
Next, suppose that $x \in A_i$ for $i \in S$, where $|S| = k$. In other words, $x$ is in only exactly $k$ of the sets $A_k$ rather than all of them, where $0 \leq k < n$.

Reminder of the inclusion/exclusion formula.

\[
|\cap_{i=1}^n A_i| = \sum_{i=1}^n |A_i| - \sum_{1 \leq i < j \leq n} |A_i \cup A_j| + \sum_{1 \leq i < j < k \leq n} |A_i \cup A_j \cup A_k| + \ldots + (-1)^{n-1}|A_1 \cup A_2 \cup \ldots \cup A_n| \tag{31.44}
\]

The l.h.s. of Equation 31.46 now contributes only 0 (zero) for this $x$. 
Next, suppose that \( x \in A_i \) for \( i \in S \), where \( |S| = k \). In other words, \( x \) is in only exactly \( k \) of the sets \( A_k \) rather than all of them, where \( 0 \leq k < n \).

Reminder of the inclusion/exclusion formula.

\[
| \bigcap_{i=1}^{n} A_i | = \sum_{i=1}^{n} |A_i| - \sum_{1 \leq i < j \leq n} |A_i \cup A_j| + \sum_{1 \leq i < j < k \leq n} |A_i \cup A_j \cup A_k| + \ldots + (-1)^{n-1}|A_1 \cup A_2 \cup \ldots \cup A_n|
\]

The l.h.s. of Equation 31.46 now contributes only 0 (zero) for this \( x \).

For the r.h.s. of Equation 31.46, then the contribution for this particular \( x \) will be . . .
For the r.h.s. of Equation 31.37, we have the contribution for this particular \( x \) be:

\[
k - \left[ \binom{n}{2} - \binom{n-k}{2} \right] + \left[ \binom{n}{3} - \binom{n-k}{3} \right] + \ldots + (-1)^{n-k-1} \left[ \binom{n}{n-k} - \binom{n-k}{n-k} \right]
\]

\[
+ (-1)^{n-k} \left[ \binom{n}{n-k+1} \right] + \ldots + (-1)^{n-1} \left[ \binom{n}{n} \right]
\]

\[
= (-1)^0 \left[ \binom{n}{1} - \binom{n-k}{1} \right] - \left[ \binom{n}{2} - \binom{n-k}{2} \right] + \left[ \binom{n}{3} - \binom{n-k}{3} \right] + \ldots + (-1)^{n-k-1} \left[ \binom{n}{n-k} - \binom{n-k}{n-k} \right]
\]

\[
+ (-1)^{n-k} \left[ \binom{n}{n-k+1} \right] + \ldots + (-1)^{n-1} \left[ \binom{n}{n} \right]
\]

\[
= (-1) \left[ (-1)^1 \binom{n}{1} + (-1)^2 \binom{n}{2} + \ldots + (-1)^n \binom{n}{n} \right] + 1 - 1
\]

\[
- \left[ (-1)^0 \binom{n-k}{1} + (-1)^1 \binom{n-k}{2} + (-1)^2 \binom{n-k}{3} + \ldots + (-1)^{n-k-1} \binom{n-k}{n-k} \right]
\]

\[
= (-1) \left[ (-1)^0 \binom{n}{0} + (-1)^1 \binom{n}{1} + (-1)^2 \binom{n}{2} + \ldots + (-1)^n \binom{n}{n} \right] - 1
\]

\[
+ \left[ (-1)^0 \binom{n-k}{0} + (-1)^1 \binom{n-k}{1} + (-1)^2 \binom{n-k}{2} + (-1)^3 \binom{n-k}{3} + \ldots + (-1)^{n-k} \binom{n-k}{n-k} \right] - 1
\]

\[
= (-1) \left[ (1 - 1)^n - 1 \right] + [(1 - 1)^{n-k} - 1] = 1 - 1 = 0
\]
The same exact argument can be used to show inclusion/exclusion formula for the signed measure $\mu$, i.e.,

$$
\mu(\bigcap_{i=1}^{n} A_i) = \sum_{i=1}^{n} \mu(A_i) - \sum_{1 \leq i < j \leq n} \mu(A_i \cup A_j)
$$

(31.52)

$$
+ \sum_{1 \leq i < j < k \leq n} \mu(A_i \cup A_j \cup A_k) + \ldots
$$

(31.53)

$$
+ (-1)^{n-1} \mu(A_1 \cup A_2 \cup \ldots \cup A_n)
$$

(31.54)
The same exact argument can be used to show inclusion/exclusion formula for the signed measure $\mu$, i.e.,

\[
\mu(\bigcap_{i=1}^{n} A_i) = \sum_{i=1}^{n} \mu(A_i) - \sum_{1 \leq i < j \leq n} \mu(A_i \cup A_j) \\
+ \sum_{1 \leq i < j < k \leq n} \mu(A_i \cup A_j \cup A_k) + \ldots \\
+ (-1)^{n-1} \mu(A_1 \cup A_2 \cup \ldots \cup A_n)
\]  

(31.52)

A “dual” form of inclusion/exclusion has the form:

\[
\mu(\bigcup_{i=1}^{n} A_i) = \sum_{i=1}^{n} \mu(A_i) - \sum_{1 \leq i < j \leq n} \mu(A_i \cap A_j) \\
+ \sum_{1 \leq i < j < k \leq n} \mu(A_i \cap A_j \cap A_k) + \ldots \\
+ (-1)^{n-1} \mu(A_1 \cap A_2 \cap \ldots \cap A_n)
\]  

(31.55)
Another (easier?, shorter) way of writing these is as:

\[
\mu(\cap_{i=1}^{n} A_i) = \sum_{k=1}^{n} (-1)^{k+1} \left( \sum_{1 \leq i_1 < i_2 < \cdots < i_k \leq n} \mu(A_{i_1} \cup \cdots \cup A_{i_k}) \right)
\]

(31.58)

and

\[
\mu(\cup_{i=1}^{n} A_i) = \sum_{k=1}^{n} (-1)^{k+1} \left( \sum_{1 \leq i_1 < i_2 < \cdots < i_k \leq n} \mu(A_{i_1} \cap \cdots \cap A_{i_k}) \right)
\]

(31.59)
What needs to be specified

**Theorem 31.3.1**

Signed measure $\mu$ on $\mathcal{F}_n$ is fully specified by $\{\mu(B) : B \in \mathcal{B}\}$ which can be any set of real numbers.

- So before, we defined $\mu$ on all atoms and noted that this defined $\mu$ everywhere else.

- Here, in the above theorem, we are defining $\mu$ only on elements of $\mathcal{B}$ and are again saying that this defines the values of $\mu$ everywhere else.

- We will see that this allows us to generate all standard mutual-information quantities, and some other less standard ones, using just entropy to fill out $\mu$ on $\mathcal{B}$.
proof of Theorem 31.3.1.

- Note that $|A| = |B| = 2^n - 1 \triangleq k$
Proof of Theorem 31.3.1

Note that $|\mathcal{A}| = |\mathcal{B}| = 2^n - 1 \triangleq k$

Define $\vec{a} = [\ldots \mu(A) \ldots]^\top$ for all $A \in \mathcal{A}$. $\text{length}(\vec{a}) = k$
Proof of Theorem 31.3.1

Note that $|A| = |B| = 2^n - 1 \triangleq k$

Define $\vec{a} = [\ldots \mu(A) \ldots]^\top$ for all $A \in \mathcal{A}$. $\text{length}(\vec{a}) = k$

Define $\vec{b} = [\ldots \mu(B) \ldots]^\top$ for all $B \in \mathcal{B}$. $\text{length}(\vec{b}) = k$
proof of Theorem 31.3.1.

- Note that $|A| = |B| = 2^n - 1 \triangleq k$
- Define $\vec{a} = [\ldots \mu(A) \ldots ]^\top$ for all $A \in A$. $\text{length}(\vec{a}) = k$
- Define $\vec{b} = [\ldots \mu(B) \ldots ]^\top$ for all $B \in B$. $\text{length}(\vec{b}) = k$
- For any $B \in B$, we have $B = \cup_{\ell \in A(B)} A_\ell$ with $A_\ell \in A$ and where $A(B)$ are the indices of the atoms that comprise $B$. 

...
proof of Theorem 31.3.1.

- Note that \(|A| = |B| = 2^n - 1 \triangleq k\)
- Define \(\vec{a} = [\ldots \mu(A) \ldots]^\top\) for all \(A \in \mathcal{A}\). \(\text{length}(\vec{a}) = k\)
- Define \(\vec{b} = [\ldots \mu(B) \ldots]^\top\) for all \(B \in \mathcal{B}\). \(\text{length}(\vec{b}) = k\)
- For any \(B \in \mathcal{B}\), we have \(B = \bigcup_{\ell \in \mathcal{A}(B)} A_{\ell}\) with \(A_{\ell} \in \mathcal{A}\) and where \(\mathcal{A}(B)\) are the indices of the atoms that comprise \(B\).
- Therefore, there exists a unique \(k \times k\) matrix \(C_n\) such that \(\vec{b} = C_n\vec{a}\)
Note that $|A| = |B| = 2^n - 1 \triangleq k$

Define $\vec{a} = [\ldots \mu(A) \ldots]^\top$ for all $A \in \mathcal{A}$. $\text{length}(\vec{a}) = k$

Define $\vec{b} = [\ldots \mu(B) \ldots]^\top$ for all $B \in \mathcal{B}$. $\text{length}(\vec{b}) = k$

For any $B \in \mathcal{B}$, we have $B = \bigcup_{\ell \in \mathcal{A}(B)} A_\ell$ with $A_\ell \in \mathcal{A}$ and where $\mathcal{A}(B)$ are the indices of the atoms that comprise $B$.

Therefore, there exists a unique $k \times k$ matrix $C_n$ such that $\vec{b} = C_n \vec{a}$

But also, we claim that any $\mu(A)$ for $A \in \mathcal{A}$ can be expressed as a linear combination of $\{\mu(B)\}_{B \in \mathcal{B}(A)}$ for the appropriate $\mathcal{B}(A) \subseteq \mathcal{B}$, and this can be done using inclusion/exclusion.

...
Here, the inclusion/exclusion principle takes the form conditioned on (or excluding) $B$:

$$
\mu(\bigcap_{k=1}^{n} \tilde{X}_k \setminus B) = \sum_{1 \leq i \leq n} \mu(\tilde{X}_i \setminus B) - \sum_{1 \leq i < j \leq n} \mu((\tilde{X}_i \cup \tilde{X}_j) \setminus B) + \cdots + (-1)^{n+1} \mu((\tilde{X}_1 \cup \tilde{X}_2 \cup \cdots \cup \tilde{X}_n) \setminus B) \tag{31.60}
$$

...
Proof of Theorem 31.3.1

Here, the inclusion/exclusion principle takes the form conditioned on (or excluding) $B$:

$$\mu(\bigcap_{k=1}^{n} \tilde{X}_k \setminus B) = \sum_{1 \leq i \leq n} \mu(\tilde{X}_i \setminus B) - \sum_{1 \leq i < j \leq n} \mu((\tilde{X}_i \cup \tilde{X}_j) \setminus B)$$

$$+ \cdots + (-1)^{n+1} \mu((\tilde{X}_1 \cup \tilde{X}_2 \cup \cdots \cup \tilde{X}_n) \setminus B)$$

(31.60)

How can this help us? Note: this formula works for any number of sets, not just all $n$ of them (so we can take $n$ to be any number of sets rather than only all of them).
Proof of Theorem 31.3.1

Since \( A \setminus B = A \cap B^c \), every atom \( A \in \mathcal{A} \) corresponds to:

\[
A = \bigcap_{i=1}^{n} Y_i = \left( \bigcap_{j:Y_j=\tilde{X}_j} \tilde{X}_j \right) \cap \left( \bigcap_{j:Y_j=\tilde{X}_j^c} \tilde{X}_j^c \right) \quad (31.61)
\]

\[
= \left( \bigcap_{j:Y_j=\tilde{X}_j} \tilde{X}_j \right) \cap \left( \bigcup_{j:Y_j=\tilde{X}_j^c} \tilde{X}_j \right) \quad (31.62)
\]

\[
= \left( \bigcup_{j:Y_j=\tilde{X}_j^c} \tilde{X}_j \right) \setminus B \quad (31.63)
\]
proof of Theorem 31.3.1.

Also, each of the terms of the r.h.s. of the inclusion/exclusion formula (Eqn.(31.60)) may take the form:

\[
\mu(\tilde{X}_i \cup \tilde{X}_j \cup \cdots \cup \tilde{X}_k \setminus B) = \mu(\tilde{X}_i \cup \tilde{X}_j \cup \cdots \cup \tilde{X}_k \setminus \bigcup_{\ell} \tilde{X}_l)
\]

(31.64)

\[
= \mu(\tilde{X}_i \cup \tilde{X}_j \cup \cdots \cup \tilde{X}_k \cup \bigcup_{\ell} \tilde{X}_l) - \mu(\bigcup_{\ell} \tilde{X}_l)
\]

(31.65)

which is true since \(\mu(U \setminus V) = \mu(U \cup V) - \mu(V)\), \(\forall U, V\).

...
Proof of Theorem 31.3.1

- Also, each of the terms of the r.h.s. of the inclusion/exclusion formula (Eqn.(31.60)) may take the form:

\[
\mu(\tilde{X}_i \cup \tilde{X}_j \cup \cdots \cup \tilde{X}_k \setminus B) = \mu(\tilde{X}_i \cup \tilde{X}_j \cup \cdots \cup \tilde{X}_k \setminus \bigcup_{\ell} \tilde{X}_\ell) = \mu(\tilde{X}_i \cup \tilde{X}_j \cup \cdots \cup \tilde{X}_k \cup \bigcup_{\ell} \tilde{X}_\ell) - \mu(\bigcup_{\ell} \tilde{X}_\ell)
\]

which is true since \(\mu(U \setminus V) = \mu(U \cup V) - \mu(V), \forall U, V\).

- Thus, the measure of any atom \(A \in \mathcal{A}\) is representable as a sum of weighted measures of the unions of the basic sets \(\mathcal{B}\)!
Proof of Theorem 31.3.1

proof of Theorem 31.3.1.

- Also, each of the terms of the r.h.s. of the inclusion/exclusion formula (Eqn.(31.60)) may take the form:

\[
\mu(\tilde{X}_i \cup \tilde{X}_j \cup \cdots \cup \tilde{X}_k \setminus B) = \mu(\tilde{X}_i \cup \tilde{X}_j \cup \cdots \cup \tilde{X}_k \setminus \bigcup_{\ell} \tilde{X}_l)
\]

(31.64)

\[
= \mu(\tilde{X}_i \cup \tilde{X}_j \cup \cdots \cup \tilde{X}_k \cup \bigcup_{\ell} \tilde{X}_l) - \mu(\bigcup_{\ell} \tilde{X}_l)
\]

(31.65)

which is true since \(\mu(U \setminus V) = \mu(U \cup V) - \mu(V)\), \(\forall U, V\).

- Thus, the measure of any atom \(A \in \mathcal{A}\) is representable as a sum of weighted measures of the unions of the basic sets \(\mathcal{B}\)!

- Therefore, there exists a \(k \times k\) matrix \(D_n\) such that \(\vec{a} = D_n \vec{b}\) (before we had \(\vec{b} = C_n \vec{a}\)).
Proof of Theorem 31.3.1

Since $C_n$ is unique, so is $D_n$ with $D_n = C_n^{-1}$.
Proof of Theorem 31.3.1

proof of Theorem 31.3.1.

- Since $C_n$ is unique, so is $D_n$ with $D_n = C_n^{-1}$.

- To summarize, we can define measure values only on $\mathcal{B}$ and it defines the measures for all elements of $\mathcal{F}_n$. 
proof of Theorem 31.3.1.

- Since $C_n$ is unique, so is $D_n$ with $D_n = C_n^{-1}$.

- To summarize, we can define measure values only on $\mathcal{B}$ and it defines the measures for all elements of $\mathcal{F}_n$.

- For example, we can define just the values $H(X_G)$ for $G \subseteq [n]$ and this defines every other information theoretic value.
Examples

- Here are two examples of the above ideas, expressible as lemmas, first proven using set theory and second proven using information theoretic ideas.

\[
\begin{align*}
\text{Let } A, B, C \text{ be sets.} \\
\text{Lemma:} \quad \mu((A \cap B) \cap C) &= \mu(A \cup C) + \mu(B \cup C) - \mu(A \cup B \cup C) - \mu(C) \\
\text{Proof:} \\
\mu((A \cap B) \cap C) &= \mu(A \cap C) + \mu(B \cap C) - \mu((A \cup B) \cap C) \\
&= \left(\mu(A \cup C) - \mu(C)\right) + \left(\mu(B \cup C) - \mu(C)\right) - \left(\mu(A \cup B \cup C) - \mu(C)\right) \\
&= \mu(A \cup C) + \mu(B \cup C) - \mu(A \cup B \cup C) - \mu(C)
\end{align*}
\]
Examples

- Here are two examples of the above ideas, expressible as lemmas, first proven using set theory and second proven using information theoretic ideas.

- Let $A, B, C$ be sets.
Here are two examples of the above ideas, expressible as lemmas, first proven using set theory and second proven using information theoretic ideas.

Let $A, B, C$ be sets.

lemma:
\[
\mu((A \cap B) \setminus C) = \mu(A \cup C) + \mu(B \cup C) - \mu(A \cup B \cup C) - \mu(C).
\]
Here are two examples of the above ideas, expressible as lemmas, first proven using set theory and second proven using information theoretic ideas.

Let $A, B, C$ be sets.

**lemma:**

$$
\mu((A \cap B) \setminus C) = \mu(A \cup C) + \mu(B \cup C) - \mu(A \cup B \cup C) - \mu(C).
$$

**proof:**

$$
\mu((A \cap B) \setminus C) = \mu(A \setminus C) + \mu(B \setminus C) - \mu((A \cup B) \setminus C) \quad (31.66)
$$

$$
= (\mu(A \cup C) - \mu(C)) + (\mu(B \cup C) - \mu(C)) - (\mu(A \cup B \cup C) - \mu(C)) \quad (31.67)
$$

$$
= \mu(A \cup C) + \mu(B \cup C) - \mu(A \cup B \cup C) - \mu(C) \quad (31.68)
$$
Examples

- As a Shannon measure (let $A, B, C$ be random variables):

\[
I(A; B | C) = H(A, C) + H(B, C) - H(A, B, C) - H(C)
\]
As a Shannon measure (let $A, B, C$ be random variables):

**Lemma:** $I(A; B|C) = H(A, C) + H(B, C) - H(A, B, C) - H(C)$
As a Shannon measure (let $A, B, C$ be random variables):

**lemma:**

$$I(A; B|C) = H(A, C) + H(B, C) - H(A, B, C) - H(C)$$

**proof:**

$$I(A; B|C) = H(A|C) - H(A|B, C)$$

$$= H(A, C) - H(C) - \left( H(A, B, C) - H(B, C) \right)$$

$$= H(A, C) + H(B, C) - H(A, B, C) + H(C)$$
As a Shannon measure (let $A, B, C$ be random variables):

**Lemma:** \[ I(A; B|C) = H(A, C) + H(B, C) - H(A, B, C) - H(C) \]

**Proof:**

\[
I(A; B|C) = H(A|C) - H(A|B, C)
= H(A, C) - H(C) - (H(A, B, C) - H(B, C)) \tag{31.69}
= H(A, C) + H(B, C) - H(A, B, C) + H(C) \tag{31.70}
\]

**Key point:** this is the same as before, but in the first case we showed it using set theory.
Another key point. The information measure generates everything. That is, assigning the measures only for $G$ as in

$$\mu(\tilde{X}_G) \triangleq H(X_G) \text{ for } G \subseteq [n]$$  \hspace{1cm} (31.72)

is sufficient to get all remaining Shannon’s information theoretic quantities.

Example: Start by defining $\mu(\tilde{X}_G)$ only on $B$ with entropic quantities as above. Then consider $G, G', G''$ as sets, and form

$$\mu((\tilde{X}_G \cap \tilde{X}_{G'})) \cap \tilde{X}_{G''}) = \mu(\tilde{X}_G \cup \tilde{X}_{G''}) + \mu(\tilde{X}_{G'} \cup \tilde{X}_{G''}) - \mu(X_G, X_{G'}, X_{G''}) - \mu(X_{G''})$$  \hspace{1cm} (31.73)

That is, in most general case, we get conditional mutual information, but setting various combinations of $G, G', G''$ to empty allows us to get $I(X_G; X_{G'} | X_{G''})$ (if $G'' = \emptyset$), $H(X_G | X_{G''})$ (if $G = G'$), or $H(X_G)$ (if $G = G'$ and $G'' = 0$).
Sufficiency

- Another key point. The information measure generates everything. That is, assigning the measures only for $G$ as in
\[
\mu(\tilde{X}_G) \triangleq H(X_G) \text{ for } G \subseteq [n]
\] (31.72)
is sufficient to get all remaining Shannon’s information theoretic quantities.

- Example: Start by defining $\mu(\tilde{X}_G)$ only on $\mathcal{B}$ with entropic quantities as above.
Another key point. The information measure generates everything. That is, assigning the measures only for $G$ as in

$$\mu(\tilde{X}_G) \triangleq H(X_G) \text{ for } G \subseteq [n]$$

(31.72)

is sufficient to get all remaining Shannon’s information theoretic quantities.

Example: Start by defining $\mu(\tilde{X}_G)$ only on $B$ with entropic quantities as above.

Then consider $G$, $G'$, and $G''$ as sets, and form

$$\mu((\tilde{X}_G \cap \tilde{X}_{G'}) \setminus \tilde{X}_{G''})$$

(31.73)

$$= \mu(\tilde{X}_G \cup \tilde{X}_{G''}) + \mu(\tilde{X}_{G'} \cup \tilde{X}_{G''}) - \mu(X_G, X_{G'}, X_{G''}) - \mu(X_{G''})$$

(31.74)

$$= I(X_G; X_{G'}|X_{G''})$$
Another key point. The information measure generates everything. That is, assigning the measures only for $G$ as in
\[
\mu(\tilde{X}_G) \triangleq H(X_G) \text{ for } G \subseteq [n]
\]  
(31.72)
is sufficient to get all remaining Shannon’s information theoretic quantities.

Example: Start by defining $\mu(\tilde{X}_G)$ only on $B$ with entropic quantities as above.

Then consider $G$, $G'$, and $G''$ as sets, and form
\[
\mu(\tilde{X}_G \cap \tilde{X}_{G'} \setminus \tilde{X}_{G''}) \\
= \mu(\tilde{X}_G \cup \tilde{X}_{G''}) + \mu(\tilde{X}_{G'} \cup \tilde{X}_{G''}) \geq \mu(X_G, X_{G'}, X_{G''}) - \mu(X_{G''}) \\
= I(X_G; X_{G'}|X_{G''}) \\
\]  
(31.73)

That is, in most general case, we get conditional mutual information, but setting various combinations of $G$, $G'$, or $G''$ to empty allows us to get $I(X_G; X_{G'})$ (if $G'' = \emptyset$), $H(X_G|X_{G''})$ (if $G' = \emptyset$), and $H(X_{G''})$ (if $G = \emptyset$).
Signed Measure

- When $n = 2$ all measures are positive given entropic instantiations (e.g., discrete entropy and MI are non-negative).
Signed Measure

- When \( n = 2 \) all measures are positive given entropic instantiations (e.g., discrete entropy and MI are non-negative).
- When \( n = 3 \), is this so?
Signed Measure

- When $n = 2$ all measures are positive given entropic instantiations (e.g., discrete entropy and MI are non-negative).
- When $n = 3$, is this so? What is:

$$\mu(\tilde{X}_1 \cap \tilde{X}_2 \cap \tilde{X}_3)$$ \hspace{1cm} (31.75)

and what might it mean?
When \( n = 2 \) all measures are positive given entropic instantiations (e.g., discrete entropy and MI are non-negative).

When \( n = 3 \), is this so? What is:

\[
\mu(\tilde{X}_1 \cap \tilde{X}_2 \cap \tilde{X}_3) \tag{31.75}
\]

and what might it mean?

We have that:

\[
\mu(\tilde{X}_1 \cap \tilde{X}_2 \cap \tilde{X}_3) + \mu(\tilde{X}_1 \cap \tilde{X}_2 \cap \tilde{X}_3^c) = \mu(\tilde{X}_1 \cap \tilde{X}_2) \tag{31.76}
\]
Signed Measure

- When $n = 2$ all measures are positive given entropic instantiations (e.g., discrete entropy and MI are non-negative).
- When $n = 3$, is this so? What is:

$$\mu(\tilde{X}_1 \cap \tilde{X}_2 \cap \tilde{X}_3)$$ (31.75)

and what might it mean?

- We have that:

$$\mu(\tilde{X}_1 \cap \tilde{X}_2 \cap \tilde{X}_3) + \mu(\tilde{X}_1 \cap \tilde{X}_2 \cap \tilde{X}_3^c) = \mu(\tilde{X}_1 \cap \tilde{X}_2)$$ (31.76)

which translates as

$$I(X_1; X_2; X_3) + I(X_1; X_2|X_3) = I(X_1; X_2)$$ (31.77)
In fact, we have that:

\[ I(X_1; X_2; X_3) = I(X_1; X_2) - I(X_1; X_2|X_3) \] (31.78)

\[ = I(X_1; X_3) - I(X_1; X_3|X_2) \] (31.79)

\[ = I(X_2; X_3) - I(X_2; X_3|X_1) \] (31.80)

Why is this?
In fact, we have that:

\[
I(X_1; X_2; X_3) = I(X_1; X_2) - I(X_1; X_2|X_3) \tag{31.78}
\]
\[
= I(X_1; X_3) - I(X_1; X_3|X_2) \tag{31.79}
\]
\[
= I(X_2; X_3) - I(X_2; X_3|X_1) \tag{31.80}
\]

Why is this? Due to the three ways of “de-eliminating” or “de-marginalizing” the three variables, that is:

\[
\mu(\tilde{X}_1 \cap \tilde{X}_2 \cap \tilde{X}_3) + \mu(\tilde{X}_1 \cap \tilde{X}_2 \cap \tilde{X}_3^c) = \mu(\tilde{X}_1 \cap \tilde{X}_2) \tag{31.81}
\]
\[
\mu(\tilde{X}_1 \cap \tilde{X}_2 \cap \tilde{X}_3) + \mu(\tilde{X}_1^c \cap \tilde{X}_2 \cap \tilde{X}_3) = \mu(\tilde{X}_1 \cap \tilde{X}_3) \tag{31.82}
\]
\[
\mu(\tilde{X}_1 \cap \tilde{X}_2 \cap \tilde{X}_3) + \mu(\tilde{X}_1 \cap \tilde{X}_2^c \cap \tilde{X}_3) = \mu(\tilde{X}_2 \cap \tilde{X}_3) \tag{31.83}
\]
Signed Measure, Information Among r.v.s

- $I(X_1; X_2; X_3)$ is our notation for this.
• $I(X_1; X_2; X_3)$ is our notation for this.
• So we have a “new” information theoretic quantity $I(X_1; X_2; X_3)$ for discrete random variables. Must it be positive?
\[ I(X_1; X_2; X_3) \] is our notation for this.

So we have a “new” information theoretic quantity \( I(X_1; X_2; X_3) \) for discrete random variables. Must it be positive?

Consider uniform binary \( X_1 \perp\!\!\!\!\!\perp X_2 \) and \( X_3 = X_1 \oplus X_2 \) (xor). Then \( X_i \perp\!\!\!\!\!\perp X_j \) for \( i \neq j \) and \( I(X_i; X_j) = 0 \) and \( H(X_i|X_j) = 1 \) for \( i \neq j \).
\( I(X_1; X_2; X_3) \) is our notation for this.

So we have a “new” information theoretic quantity \( I(X_1; X_2; X_3) \) for discrete random variables. Must it be positive?

Consider uniform binary \( X_1 \perp \perp X_2 \) and \( X_3 = X_1 \oplus X_2 \) (xor). Then \( X_i \perp \perp X_j \) for \( i \neq j \) and \( I(X_i; X_j) = 0 \) and \( H(X_i|X_j) = 1 \) for \( i \neq j \)

Furthermore, \( H(X_i|X_j, X_k) = 0 \) for \( i \neq j, j \neq k, k \neq i \), and \( I(X_1; X_2|X_3) = H(X_1|X_3) - H(X_1|X_3, X_2) = 1 - 0 = 1. \)
$I(X_1; X_2; X_3)$ is our notation for this.

So we have a “new” information theoretic quantity $I(X_1; X_2; X_3)$ for discrete random variables. Must it be positive?

Consider uniform binary $X_1 \perp \perp X_2$ and $X_3 = X_1 \oplus X_2$ (xor). Then $X_i \perp \perp X_j$ for $i \neq j$ and $I(X_i; X_j) = 0$ and $H(X_i|X_j) = 1$ for $i \neq j$.

Furthermore, $H(X_i|X_j, X_k) = 0$ for $i \neq j$, $j \neq k$, $k \neq i$, and $I(X_1; X_2|X_3) = H(X_1|X_3) - H(X_1|X_3, X_2) = 1 - 0 = 1$.

Therefore,

$I(X_1; X_2; X_3) = I(X_1; X_2) - I(X_1; X_2|X_3) = 0 - 1 = -1$!
More on $I(X_1; X_2; X_3)$

This above is a simple consequence of that conditioning can either increase or decrease mutual information.
This above is a simple consequence of that conditioning can either increase or decrease mutual information.

Example: Increase, example from before $X_3 = X_1 \oplus X_2$ with $0 = I(X_1; X_2) < I(X_1; X_2|X_3) = 1$
More on $I(X_1; X_2; X_3)$

- This above is a simple consequence of that conditioning can either increase or decrease mutual information.

- Example: Increase, example from before $X_3 = X_1 \oplus X_2$ with $0 = I(X_1; X_2) < I(X_1; X_2|X_3) = 1$

- Example: Decrease, Markov chain $X_1 \rightarrow X_2 \rightarrow X_3$ so $I(X_1; X_3) \geq I(X_1; X_3|X_2) = 0$. 
More on $I(X_1; X_2; X_3)$

- This above is a simple consequence of that conditioning can either increase or decrease mutual information.
- Example: Increase, example from before $X_3 = X_1 \oplus X_2$ with $0 = I(X_1; X_2) < I(X_1; X_2|X_3) = 1$
- Example: Decrease, Markov chain $X_1 \rightarrow X_2 \rightarrow X_3$ so $I(X_1; X_3) \geq I(X_1; X_3|X_2) = 0$.
- Thus $I(X_1; X_2; X_3)$ might be either positive or negative.
More on $I(X_1; X_2; X_3)$

- We’ve seen $I(X_1; X_2; X_3)$ before in Venn diagrams.
More on $I(X_1; X_2; X_3)$

- We’ve seen $I(X_1; X_2; X_3)$ before in Venn diagrams.
- We can thus view $I(X_1; X_2; X_3)$ as a form of information “among” three random variables, or the information “common to” the three r.v.s.
In fact,

\[
I(X_1; X_2; X_3) = I(X_1; X_2) - I(X_1; X_2|X_3) \tag{31.84}
\]

\[
= I(X_1; X_2) - [I(X_1; X_2, X_3) - I(X_1; X_3)] \tag{31.85}
\]

\[
= [I(X_1; X_2) + I(X_1; X_3)] - I(X_1; X_2, X_3) \tag{31.86}
\]
More on $I(X_1; X_2; X_3)$

- In fact,

\[
I(X_1; X_2; X_3) = I(X_1; X_2) - I(X_1; X_2|X_3) \tag{31.84}
\]

\[
= I(X_1; X_2) - [I(X_1; X_2, X_3) - I(X_1; X_3)] \tag{31.85}
\]

\[
= [I(X_1; X_2) + I(X_1; X_3)] - I(X_1; X_2, X_3) \tag{31.86}
\]

- So $I(X_1; X_2; X_3)$ can be seen as the difference between the joint information $X_2, X_3$ has about $X_1$, and the sum of the individual amounts of information $X_2$ and $X_3$ each have about $X_1$. 
More on $I(X_1; X_2; X_3)$

- In fact,

\[
I(X_1; X_2; X_3) = I(X_1; X_2) - I(X_1; X_2 | X_3) \tag{31.84}
\]

\[
= I(X_1; X_2) - [I(X_1; X_2, X_3) - I(X_1; X_3)] \tag{31.85}
\]

\[
= [I(X_1; X_2) + I(X_1; X_3)] - I(X_1; X_2, X_3) \tag{31.86}
\]

- So $I(X_1; X_2; X_3)$ can be seen as the difference between the joint information $X_2, X_3$ has about $X_1$, and the sum of the individual amounts of information $X_2$ and $X_3$ each have about $X_1$.

- Of course, this is symmetric so this could be regarding the information between any two $X_i, X_j$ and a third $X_k$ for $i \neq j$, $j \neq k$, and $k \neq i$. 
An alternate to $I(X_1; X_2; X_3)$

As an alternative (note the prime), consider:

$$I'(X_1; X_2; X_3) \triangleq H(X_1) + H(X_2) + H(X_3) - H(X_1, X_2, X_3)$$

(31.87)

$$= \sum_{x_1, x_2, x_3} p(x_1, x_2, x_3) \log \frac{p(x_1, x_2, x_3)}{p(x_1)p(x_2)p(x_3)}$$

(31.88)
An alternate to $I(X_1; X_2; X_3)$

- As an alternative (note the prime), consider:

\[ I'(X_1; X_2; X_3) \triangleq H(X_1) + H(X_2) + H(X_3) - H(X_1, X_2, X_3) \]  

(31.87)

\[ = \sum_{x_1, x_2, x_3} p(x_1, x_2, x_3) \log \frac{p(x_1, x_2, x_3)}{p(x_1)p(x_2)p(x_3)} \]  

(31.88)

- So $I'(X_1; X_2; X_3)$ is the difference in coding length between coding independently and coding jointly (which is 0 if all variables are independent, and large if the variables are highly correlated).
An alternate to $I(X_1; X_2; X_3)$

- As an alternative (note the prime), consider:

$$I'(X_1; X_2; X_3) \triangleq H(X_1) + H(X_2) + H(X_3) - H(X_1, X_2, X_3)$$ (31.89)
An alternate to $I(X_1; X_2; X_3)$

As an alternative (note the prime), consider:

$$I'(X_1; X_2; X_3) \triangleq H(X_1) + H(X_2) + H(X_3) - H(X_1, X_2, X_3)$$

(31.89)

Visualized, we have:
Applications of $I(X_1; X_2; X_3)$

- **EAR measure** in pattern classification:
  
  \[-I(X_i; X_j; C) = I(X_i; X_j | C) - I(X_i; X_j)\]

  where $C$ is a class variable and $X_i$ are features. If EAR is positive, then $X_i, X_j$ are more dependent conditioned on $C$ than otherwise (so is indicative of a “good” direct interaction to model for classification).

- Synergy in a nerve cell (neural code of sensory stimuli). Here, $S =$ stimuli, $R_1 =$ response of one neuron, $R_2 =$ response of another neuron. Then “synergy ($R_1, R_2$)" is defined as:

  \[\text{synergy} (R_1, R_2) = I(S; R_1, R_2) - I(S; R_1) - I(S; R_2)\]

  \[= I(R_1; R_2 | S) - I(R_1; R_2)\]

  \[= -I(R_1; R_2; S)\]

  If synergy < 0, there exists “redundancy” in the neural code. If synergy > 0, the cells are “synergistic”, they encode more information in their neural response about a stimuli jointly than the sum of what they cover separately.
**Applications of $I(X_1; X_2; X_3)$**

- **EAR measure** in pattern classification:
  \[-I(X_i; X_j; C) = I(X_i; X_j|C) - I(X_i; X_j)\]
  where $C$ is a class variable and $X_i$ are features. If EAR is positive, then $X_i, X_j$ are more dependent conditioned on $C$ than otherwise (so is indicative of a “good” direct interaction to model for classification).

- **Synergy** in a nerve cell (neural code of sensory stimuli). Here, $S =$ stimuli, $R_1 =$ response of one neuron, $R_2 =$ response of another neuron. Then “synergy($R_1, R_2$)” is defined as:
  \[
  \text{synergy}(R_1, R_2) = I(S; R_1, R_2) - I(S; R_1) - I(S; R_2) \tag{31.90}
  
  = I(R_1; R_2|S) - I(R_1; R_2) \tag{31.91}
  
  = -I(R_1; R_2; S) \tag{31.92}
  \]
Applications of $I(X_1; X_2; X_3)$

- **EAR measure** in pattern classification:
  
  $$-I(X_i; X_j; C) = I(X_i; X_j|C) - I(X_i; X_j)$$
  
  where $C$ is a class variable and $X_i$ are features. If EAR is positive, then $X_i, X_j$ are more dependent conditioned on $C$ than otherwise (so is indicative of a “good” direct interaction to model for classification).

- **Synergy** in a nerve cell (neural code of sensory stimuli). Here, $S =$ stimuli, $R_1 =$ response of one neuron, $R_2 =$ response of another neuron. Then “synergy$(R_1, R_2)$” is defined as:

  $$\text{synergy}(R_1, R_2) = I(S; R_1, R_2) - I(S; R_1) - I(S; R_2)$$  \hspace{1cm} (31.90)

  $$= I(R_1; R_2|S) - I(R_1; R_2)$$  \hspace{1cm} (31.91)

  $$= -I(R_1; R_2; S)$$  \hspace{1cm} (31.92)

- If synergy $< 0$, there exists “redundancy” in the neural code.
Applications of $I(X_1; X_2; X_3)$

- **EAR measure** in pattern classification:
  $$-I(X_i; X_j; C) = I(X_i; X_j|C) - I(X_i; X_j)$$
  where $C$ is a class variable and $X_i$ are features. If EAR is positive, then $X_i$, $X_j$ are more dependent conditioned on $C$ than otherwise (so is indicative of a “good” direct interaction to model for classification).

- **Synergy** in a nerve cell (neural code of sensory stimuli). Here, $S =$ stimuli, $R_1 =$ response of one neuron, $R_2 =$ response of another neuron. Then “synergy($R_1, R_2$)” is defined as:
  $$\text{synergy}(R_1, R_2) = I(S; R_1, R_2) - I(S; R_1) - I(S; R_2) \quad (31.90)$$
  $$= I(R_1; R_2|S) - I(R_1; R_2) \quad (31.91)$$
  $$= -I(R_1; R_2; S) \quad (31.92)$$

- If synergy < 0, there exists “redundancy” in the neural code.
- If synergy > 0, there the cells are “synergistic”, they encode more information in their neural response about a stimuli jointly than the sum of what they cover separately.
For $n > 3$ random variables, and multiple groups $\{G_i\}_i$, we can define:

$$\mu((\tilde{X}_{G_1} \cap \tilde{X}_{G_2} \cap \cdots \cap \tilde{X}_{G_m}) \setminus \tilde{X}_F) \triangleq I(X_{G_1}; X_{G_2}; \ldots; X_{G_m} | X_F)$$

(31.93)

as the mutual information “between” or “among” the groups $\{X_{G_i}\}_i$ conditioned on $X_F$
Example: Two views of two views of a source

- Can correlated noisy signals ever reveal more about a source than independent (noisy) looks at a source?
Example: Two views of two views of a source

- Can correlated noisy signals ever reveal more about a source than independent (noisy) looks at a source?
- We wish to learn about $X$ and we have two ways we might do it:

```
         X
        / \  \\
       /   \  \\
      /     \  \\
     /       \  \\
    Y_1      Y_2  \\
    |        |    \\
   Z_1  ---Z_2--- \\
```

vs.

```
         X
        /   \\
       /     \\
      /       \\
     /         \\
    Y_1      --- W_1 --- W_2
```

Ex: communications: might get "stuck" with noise given in $Y_1$.

Example: population theory. When testing about the instance, say, of a disease in a population, we can either test one individual twice (each might have errors) or two individuals once. Which is better?
Example: Two views of two views of a source

- Can correlated noisy signals ever reveal more about a source than independent (noisy) looks at a source?
- We wish to learn about $X$ and we have two ways we might do it:

\[
\begin{align*}
X & \quad \text{vs.} \\
Y_1 & \quad Y_2 \\
Z_1 & \quad Z_2 \\
W_1 & \quad W_2
\end{align*}
\]

- I.e., you have either $(Z_1, Z_2)$ or $(W_1, W_2)$ but not both, where $I(Z_1; X) = I(Z_2; X) = I(W_1; X) = I(W_2; X)$, so each individual variable is equally useful to learn about $X$. 

Ex: communications: might get "stuck" with noise given in $Y_1$.

Example: population theory. When testing about the instance, say, of a disease in a population, we can either test one individual twice (each might have errors) or two individuals once. Which is better?
Example: Two views of two views of a source

- Can correlated noisy signals ever reveal more about a source than independent (noisy) looks at a source?
- We wish to learn about $X$ and we have two ways we might do it:

$$X \xleftarrow{} Y_1 \xrightarrow{} Z_1 \quad \text{vs.} \quad X \xrightarrow{} Y_1 \xleftarrow{} W_1$$

- I.e., you have either $(Z_1, Z_2)$ or $(W_1, W_2)$ but not both, where $I(Z_1; X) = I(Z_2; X) = I(W_1; X) = I(W_2; X)$, so each individual variable is equally useful to learn about $X$.
- Ex: communications: might get “stuck” with noise given in $Y_1$. 
Example: Two views of two views of a source

- Can correlated noisy signals ever reveal more about a source than independent (noisy) looks at a source?
- We wish to learn about $X$ and we have two ways we might do it:

\[
\begin{array}{c}
\text{\hspace{1cm} Y}_1 \quad \text{\hspace{1cm} Y}_2 \\
\hspace{2cm} Z_1 \quad \hspace{2cm} Z_2
\end{array}
\]

vs.

\[
\begin{array}{c}
\text{\hspace{1cm} Y}_1 \\
\hspace{2cm} W_1 \quad \hspace{2cm} W_2
\end{array}
\]

- I.e., you have either $(Z_1, Z_2)$ or $(W_1, W_2)$ but not both, where $I(Z_1; X) = I(Z_2; X) = I(W_1; X) = I(W_2; X)$, so each individual variable is equally useful to learn about $X$.
- Ex: communications: might get “stuck” with noise given in $Y_1$.
- Example: population theory. When testing about the instance, say, of a disease in a population, we can either test one individual twice (each might have errors) or two individuals once. Which is better?
Example: Two views of two views of a source

- Since each variable is individually equally informative, correlated views of a source being better entirely depends on if
  \[ I(X; Z_2|Z_1) < I(X; W_2|W_1) \]
  or not.

Thus, a sufficient condition for this to be true is:

\[ I(X; W_2) < I(X; W_2|W_1) \] (31.95)

or

\[ I(X; W_2) - I(X; W_2|W_1) = I(X; W_2; W_1) < 0 \] (31.96)
Example: Two views of two views of a source

Since each variable is individually equally informative, correlated views of a source being better entirely depends on if $I(X; Z_2 | Z_1) < I(X; W_2 | W_1)$ or not.

But $Z_1 \perp \perp Z_2 | X$ so $Z_1 \rightarrow X \rightarrow Z_2$ meaning

$$I(X; Z_2 | Z_1) \leq I(X; Z_2) = I(X; W_2)$$ \hspace{1cm} (31.94)
Example: Two views of two views of a source

Since each variable is individually equally informative, correlated views of a source being better entirely depends on if $I(X; Z_2 | Z_1) < I(X; W_2 | W_1)$ or not.

But $Z_1 \perp \perp Z_2 | X$ so $Z_1 \rightarrow X \rightarrow Z_2$ meaning

$$I(X; Z_2 | Z_1) \leq I(X; Z_2) = I(X; W_2) \quad (31.94)$$

Thus, a sufficient condition for this to be true is:

$$I(X; W) < I(X; W_2 | W_1) \quad (31.95)$$

or

$$I(X; W_2) - I(X; W_2 | W_1) = I(X; W_2; W_1) < 0 \quad (31.96)$$
Example: Two views of two views of a source

- Since each variable is individually equally informative, correlated views of a source being better entirely depends on if \( I(X; Z_2|Z_1) < I(X; W_2|W_1) \) or not.
- But \( Z_1 \perp \perp Z_2|X \) so \( Z_1 \rightarrow X \rightarrow Z_2 \) meaning

\[
I(X; Z_2|Z_1) \leq I(X; Z_2) = I(X; W_2) \quad (31.94)
\]

- Thus, a sufficient condition for this to be true is:

\[
I(X; W) < I(X; W_2|W_1) \quad (31.95)
\]

or

\[
I(X; W_2) - I(X; W_2|W_1) = I(X; W_2; W_1) < 0 \quad (31.96)
\]

- Exercise: find an example of this where \( X \rightarrow Y \) are Z-channels, and \( Y \rightarrow Z \) and \( Y \rightarrow W \) are BSCs.
How far can pictures go? We saw earlier that for $n = 3$, we can show all quantities. But if $I(X_1; X_2; X_3) < 0$ then we need a way of showing “negative area”, and one way would be to do something like:

\[
\begin{array}{c}
\text{0} \\
\text{1} \\
\text{1} \\
\text{0} \\
\text{1} \\
\text{0} \\
\end{array}
\]

For $n > 3$, it is not possible to perfectly display such a diagram in 2D since for $n$ random variables we need $n-1$ dimensions to be displayed perfectly. “Perfectly” means that atoms are adjacent if, say, only one variable is complemented between the atoms. For example, \( \tilde{X}_1 \cap \tilde{X}_2 \cap \tilde{X}_3 \cap \tilde{X}_4 \) should be adjacent to \( \tilde{X}_1 \cap \tilde{X}_2 \cap \tilde{X}_3 \cap \tilde{X}_4 \).
How far can pictures go? We saw earlier that for $n = 3$, we can show all quantities. But if $I(X_1; X_2; X_3) < 0$ then we need a way of showing “negative area”, and one way would be to do something like:

For $n > 3$, it is not possible to perfectly display such a diagram in 2D since for $n$ random variables we need $n - 1$ dimensions to be displayed perfectly.
How far can pictures go? We saw earlier that for $n = 3$, we can show all quantities. But if $I(X_1; X_2; X_3) < 0$ then we need a way of showing “negative area”, and one way would be to do something like:

For $n > 3$, it is not possible to perfectly display such a diagram in 2D since for $n$ random variables we need $n − 1$ dimensions to be displayed perfectly.

“perfectly” means that atoms are adjacent if, say, only one variable is complemented between the atoms. For example, $\tilde{X}_1 \cap \tilde{X}_2 \cap \tilde{X}_3 \cap \tilde{X}_4^c$ should be adjacent to $\tilde{X}_1 \cap \tilde{X}_2 \cap \tilde{X}_3^c \cap \tilde{X}_4^c$.
One way to try to display this, almost perfectly, is as follows:
One way to try to display this, almost perfectly, is as follows:
One way to try to display this, almost perfectly, is as follows:

How many atoms should there be?
One way to try to display this, almost perfectly, is as follows:

How many atoms should there be? $2^n - 1 = 15$ since $n = 4$ here. We find 15 by counting as well.
One way to try to display this, almost perfectly, is as follows:

How many atoms should there be?
\[2^n - 1 = 15\] since \(n = 4\) here. We find 15 by counting as well.

Here we see that the set \(\tilde{X}_1 \cap \tilde{X}_2 \cap \tilde{X}_4^c\) is partitioned into \(\tilde{X}_1 \cap \tilde{X}_2 \cap \tilde{X}_3 \cap \tilde{X}_4^c\) and \(\tilde{X}_1 \cap \tilde{X}_2 \cap \tilde{X}_3^c \cap \tilde{X}_4^c\) which are not adjacent.
Actually, the measures can take any non-negative value on the atoms due to the flexibility of entropy for independent random variables. That is, we have

\[ \text{Theorem 31.3.2} \]

If there is no constraint on random variables \( X_1, X_2, \ldots, X_n \), then \( \mu^* \) can take on any set of nonnegative values on the non-empty atoms of \( F_n \)(namely \( A \)).

proof sketch.

We associate the atoms with a set of mutually independent random variables. Any random variable \( X_i \) than has entropy equal to the sum of the atoms contained in \( X_i \)’s set (i.e., \( \tilde{X}_i \)) but these individual entropies are unrestricted.
Actually, the measures can take any non-negative value on the atoms due to the flexibility of entropy for independent random variables. That is, we have

**Theorem 31.3.2**

*If there is no constraint on random variables $X_1, X_2, \ldots, X_n$, then $\mu^*$ can take on any set of nonnegative values on the non-empty atoms of $\mathcal{F}_n$ (namely $A$).*
Actually, the measures can take any non-negative value on the atoms due to the flexibility of entropy for independent random variables. That is, we have

**Theorem 31.3.2**

*If there is no constraint on random variables $X_1, X_2, \ldots, X_n$, then $\mu^*$ can take on any set of nonnegative values on the non-empty atoms of $\mathcal{F}_n$ (namely $\mathcal{A}$).*

**proof sketch.**

We associate the atoms with a set of mutually independent random variables. Any random variable $X_i$ than has entropy equal to the sum of the atoms contained in $X_i$’s set (i.e., $\tilde{X}_i$) but these individual entropies are unrestricted.
What if there are restrictions? I.e., what if we have a Markov chain $X_1 \rightarrow X_2 \rightarrow \cdots \rightarrow X_n$. 

Ex. For $n=3$, we have $X_1 \rightarrow X_2 \rightarrow X_3$ which means that

$$I(X_1;X_3|X_2)=\mu(\tilde{X}_1 \cap \tilde{X}_c \cap \tilde{X}_3)=0 \quad (31.97)$$

One way of plotting a Venn-like diagram is as follows:
What if there are restrictions? I.e., what if we have a Markov chain $X_1 \rightarrow X_2 \rightarrow \cdots \rightarrow X_n$.

Ex. For $n = 3$, we have $X_1 \rightarrow X_2 \rightarrow X_3$ which means that

$$I(X_1; X_3|X_2) = \mu(\tilde{X}_1 \cap \tilde{X}_2^c \cap \tilde{X}_3) = 0$$

(31.97)
What if there are restrictions? I.e., what if we have a Markov chain $X_1 \rightarrow X_2 \rightarrow \cdots \rightarrow X_n$.

Ex. For $n = 3$, we have $X_1 \rightarrow X_2 \rightarrow X_3$ which means that

$$I(X_1; X_3|X_2) = \mu(\tilde{X}_1 \cap \tilde{X}_2^c \cap \tilde{X}_3) = 0$$

One way of plotting a Venn-like diagram is as follows:
In this case, what happens to \( I(X_1; X_2; X_3) \)?

\[
I(X_1; X_2; X_3) = \mu(\tilde{X}_1 \cap \tilde{X}_2 \cap \tilde{X}_3) = \mu(\tilde{X}_1 \cap \tilde{X}_3) = I(X_1; X_3) \geq 0
\]
In this case, what happens to $I(X_1; X_2; X_3)$?

\[
I(X_1; X_2; X_3) = \mu(\tilde{X}_1 \cap \tilde{X}_2 \cap \tilde{X}_3) = \mu(\tilde{X}_1 \cap \tilde{X}_3) = I(X_1; X_3) \geq 0
\] (31.98)

So with a Markov chain, we have $I(X_1; X_2; X_3) \geq 0$. Another Venn-like figure would show this:

\[
\begin{array}{c}
\tilde{X}_1 \\
\tilde{X}_2 \\
\tilde{X}_3
\end{array}
\]

but it does not nicely generalize to $n > 3$. 
With larger $n$, we can deduce that certain sets are empty.
Markov Chain Restrictions

- With larger $n$, we can deduce that certain sets are empty.
- For example with $n = 4$ and $X_1 \rightarrow X_2 \rightarrow X_3 \rightarrow X_4$, we have:

$$
\mu(\tilde{X}_1 \cap \tilde{X}_3 \cap \tilde{X}_4 \cap \tilde{X}_2^c) + \mu(\tilde{X}_1 \cap \tilde{X}_3 \cap \tilde{X}_4^c \cap \tilde{X}_2^c) = \mu(\tilde{X}_1 \cap \tilde{X}_3 \cap \tilde{X}_2^c)
$$
With larger $n$, we can deduce that certain sets are empty.

For example with $n = 4$ and $X_1 \rightarrow X_2 \rightarrow X_3 \rightarrow X_4$, we have:

$$
\mu(\tilde{X}_1 \cap \tilde{X}_3 \cap \tilde{X}_4 \cap \tilde{X}_2^c) + \mu(\tilde{X}_1 \cap \tilde{X}_3 \cap \tilde{X}_4^c \cap \tilde{X}_2^c) = \mu(\tilde{X}_1 \cap \tilde{X}_3 \cap \tilde{X}_2^c)
$$

which means that

$$
I(X_1; X_3; X_4|X_2) + I(X_1; X_3|X_2, X_4) = I(X_1; X_3|X_2) = 0
$$

(31.99)
Markov Chain Restrictions

- With larger $n$, we can deduce that certain sets are empty.
- For example with $n = 4$ and $X_1 \rightarrow X_2 \rightarrow X_3 \rightarrow X_4$, we have:

$$\mu(\tilde{X}_1 \cap \tilde{X}_3 \cap \tilde{X}_4 \cap \tilde{X}_2^c) + \mu(\tilde{X}_1 \cap \tilde{X}_3 \cap \tilde{X}_4^c \cap \tilde{X}_2^c) = \mu(\tilde{X}_1 \cap \tilde{X}_3 \cap \tilde{X}_2^c)$$

which means that

$$I(X_1; X_3; X_4|X_2) + I(X_1; X_3|X_2, X_4) = I(X_1; X_3|X_2) = 0$$

(31.99)

- We can see this from the diagram:
In fact, 5 atoms have empty measures.

\[ A = H(X_1|X_2, X_3, X_4), \quad B = I(X_1; X_2|X_3, X_4), \quad C = I(X_1; X_3|X_4), \ldots, \quad \text{and} \quad G = I(X_2; X_4|X_1). \]
Markov Chain Restrictions: Empty measures

- In fact, 5 atoms have empty measures.

\[
A = H(X_1 | X_2, X_3, X_4), \quad B = I(X_1 ; X_2 | X_3, X_4),
\]
\[
C = I(X_1 ; X_3 | X_4), \ldots, \quad \text{and} \quad G = I(X_2 ; X_4 | X_1).
\]

- For example: \( I(X_1 ; X_3 | X_2) = 0 \). What else?
In fact, 5 atoms have empty measures.  

Then \( A = H(X_1|X_2, X_3, X_4) \), \( B = I(X_1; X_2|X_3, X_4) \), \( C = I(X_1; X_3|X_4) \), \ldots, and \( G = I(X_2; X_4|X_1) \).

For example: \( I(X_1; X_3|X_2) = 0 \). What else? \( I(X_1; X_4|X_2) = 0 \), \( I(X_1; X_4|X_3) = 0 \), \( I(X_2; X_4|X_3) = 0 \), and \( I(X_1, X_2; X_4|X_3) = 0 \).
In fact, 5 atoms have empty measures.

\[
A = H(X_1|X_2, X_3, X_4), \quad B = I(X_1; X_2|X_3, X_4),
C = I(X_1; X_3|X_4), \ldots, \quad \text{and} \quad G = I(X_2; X_4|X_1).
\]

For example: \(I(X_1; X_3|X_2) = 0\). What else? \(I(X_1; X_4|X_2) = 0\), \(I(X_1; X_4|X_3) = 0\), \(I(X_2; X_4|X_3) = 0\), and \(I(X_1, X_2; X_4|X_3) = 0\).

The 10 non-empty atoms are then:

\[
H(X_1|X_2, X_3, X_4),
I(X_1; X_2|X_3, X_4), \quad I(X_1; X_3|X_4), \quad I(X_1; X_4), \quad H(X_2|X_1, X_2, X_4),
I(X_2; X_3|X_1, X_4), \quad I(X_2; X_4|X_1), \quad H(X_3|X_1, X_2, X_4),
I(X_3; X_4|X_1, X_2), \quad \text{and} \quad H(X_4|X_1, X_2, X_3).
\]
Theorem 31.3.3

Given \( X_1 \rightarrow X_2 \rightarrow \cdots \rightarrow X_n \), then \( \mu() \) is always non-negative.

- It is also possible, using \( \mu \) to prove the concavity of entropy, convexity of MI in \( p(y|x) \) for fixed \( p(x) \) and the concavity of MI in \( p(x) \) for fixed \( p(y|x) \).
Theorem 31.3.3

Given $X_1 \rightarrow X_2 \rightarrow \cdots \rightarrow X_n$, then $\mu()$ is always non-negative.

- It is also possible, using $\mu$ to prove the concavity of entropy, convexity of MI in $p(y|x)$ for fixed $p(x)$ and the concavity of MI in $p(x)$ for fixed $p(y|x)$.

- Aside: perhaps an easy way to remember this: 1) remember that $H(X)$ is concave in $p(x)$ via the plot; 2) since

$$I(X;Y) = H(X) - H(X|Y) = H(X) - \sum_{x,y} p(x)p(y|x) \log p(y|x)$$

then fixing $p(x)$, we have a “convex-looking” function of $p(y|x)$ (not really though since $-p \log p$ is concave in $p$), and fixing $p(y|x)$ we have a concave-looking function of $p(x)$. 

Imperfect Secrecy Theorem

Suppose that $X$ is plain text, $Y$ is cipher text, and $Z$ is a key (password) in the following model:

$$X \xrightarrow{Z} Y$$

Goal: We need to make $I(X; Y)$ small (ideally zero). Also, need to have $H(X | Y, Z) = 0$ to be able to decrypt without error (i.e., $X \rightarrow Y \leftarrow Z$).

Under this constraint, however, we have lower bound:

$$I(X; Y) \geq H(X) - H(Z)$$

Thus, to make $I(X; Y)$ small, need to make $H(Z)$ big (long random passwords, best is to have password as long or as random as the cleartext!).

We can use information measures to show this.
Suppose that $X$ is plain text, $Y$ is cipher text, and $Z$ is a key (password) in the following model:

Goal: We need to make $I(X; Y)$ small (ideally zero).
Imperfect Secrecy Theorem

- Suppose that $X$ is plain text, $Y$ is cipher text, and $Z$ is a key (password) in the following model:

- Goal: We need to make $I(X; Y)$ small (ideally zero).

- Also, need to have $H(X|Y, Z) = 0$ to be able to decrypt without error (i.e., $X \rightarrow Y \leftarrow Z$).
Suppose that $X$ is plain text, $Y$ is cipher text, and $Z$ is a key (password) in the following model:

Goal: We need to make $I(X; Y)$ small (ideally zero).

Also, need to have $H(X|Y, Z) = 0$ to be able to decrypt without error (i.e., $X \rightarrow Y \leftarrow Z$).

Under this constraint, however, we have lower bound:

$$I(X; Y) \geq H(X) - H(Z)$$

(31.100)

thus, to make $I(X; Y)$ small, need to make $H(Z)$ big (long random passwords, best is to have password as long or as random as the cleartext!).
Suppose that $X$ is plain text, $Y$ is cipher text, and $Z$ is a key (password) in the following model:

Goal: We need to make $I(X; Y)$ small (ideally zero). Also, need to have $H(X|Y, Z) = 0$ to be able to decrypt without error (i.e., $X \rightarrow Y \leftarrow Z$).

Under this constraint, however, we have lower bound:

$$I(X; Y) \geq H(X) - H(Z)$$

(31.100)

thus, to make $I(X; Y)$ small, need to make $H(Z)$ big (long random passwords, best is to have password as long or as random as the cleartext!).

We can use information measures to show this.
Imperfect Secrecy Theorem

Consider the next info diagram

\[ I(X; Y | Z) = a \geq 0 \]
\[ I(Y; Z | X) = b \geq 0 \]
\[ H(Z | X, Y) = c \geq 0 \]
\[ I(X; Y; Z) = d \]

Then we have
\[ 0 \leq I(Y; Z) = \mu(Y \cap Z) = \mu(Y \cap Z \cap X) + \mu(Y \cap Z \cap \neg X) \]
\[ = I(X; Y; Z) + I(Y; Z | X) = b + d \]

Thus, \[ d \geq -b \]

\[ H(X) - H(Z) = (a + d) - (d + b + c) = a - b - c \]

so

\[ I(X; Y) = a + d \geq a - b \geq a - b - c = H(X) - H(Z) \]
Imperfect Secrecy Theorem

Consider the next info diagram

\[ \begin{align*}
I(X; Y|Z) &= a \geq 0 \\
I(Y; Z|X) &= b \geq 0 \\
H(Z|X, Y) &= c \geq 0 \\
I(X; Y; Z) &= d
\end{align*} \]
Consider the next info diagram

\[I(X; Y | Z) = a \geq 0\]
\[I(Y; Z | X) = b \geq 0\]
\[H(Z | X, Y) = c \geq 0\]
\[I(X; Y; Z) = d\]

Thus, \(d \geq -b\), and
\[H(X) - H(Z) = (a + d) - (d + b + c) = a - b - c\]
Imperfect Secrecy Theorem

Consider the next info diagram

\[ I(X; Y | Z) = a \geq 0 \]
\[ I(Y; Z | X) = b \geq 0 \]
\[ H(Z | X, Y) = c \geq 0 \]
\[ I(X; Y; Z) = d \]

Then we have

\[ 0 \]
Imperfect Secrecy Theorem

Consider the next info diagram

\[ I(X; Y|Z) = a \geq 0 \]
\[ I(Y; Z|X) = b \geq 0 \]
\[ H(Z|X, Y) = c \geq 0 \]
\[ I(X; Y; Z) = d \]

Then we have

\[ 0 \leq I(Y; Z) \]
Imperfect Secrecy Theorem

Consider the next info diagram

\[
\begin{align*}
I(X; Y | Z) &= a \geq 0 \\
I(Y; Z | X) &= b \geq 0 \\
H(Z | X, Y) &= c \geq 0 \\
I(X; Y; Z) &= d
\end{align*}
\]

Then we have

\[
0 \leq I(Y; Z) = \mu(Y \cap Z)
\]

(31.102)
Imperfect Secrecy Theorem

Consider the next info diagram

\[
\begin{align*}
I(X; Y|Z) &= a \geq 0 \\
I(Y; Z|X) &= b \geq 0 \\
H(Z|X, Y) &= c \geq 0 \\
I(X; Y; Z) &= d
\end{align*}
\]

Then we have

\[
0 \leq I(Y; Z) = \mu(Y \cap Z) = \mu(Y \cap Z \cap X) + \mu(Y \cap Z \cap X^c)
\]

(31.101)

(31.102)
Imperfect Secrecy Theorem

Consider the next info diagram

\[ I(X; Y|Z) = a \geq 0 \]
\[ I(Y; Z|X) = b \geq 0 \]
\[ H(Z|X, Y) = c \geq 0 \]
\[ I(X; Y; Z) = d \]

Then we have

\[ 0 \leq I(Y; Z) = \mu(Y \cap Z) = \mu(Y \cap Z \cap X) + \mu(Y \cap Z \cap X^c) \]  \hspace{1cm} (31.101)

\[ = I(X; Y; Z) + I(Y; Z|X) \]  \hspace{1cm} (31.102)
Imperfect Secrecy Theorem

Consider the next info diagram

\[ I(X; Y|Z) = a \geq 0 \]
\[ I(Y; Z|X) = b \geq 0 \]
\[ H(Z|X, Y) = c \geq 0 \]
\[ I(X; Y; Z) = d \]

Then we have

\[ 0 \leq I(Y; Z) = \mu(Y \cap Z) = \mu(Y \cap Z \cap X) + \mu(Y \cap Z \cap X^c) \] (31.101)

\[ = I(X; Y; Z) + I(Y; Z|X) = b + d \] (31.102)
Imperfect Secrecy Theorem

- Consider the next info diagram

\[
\begin{align*}
0 & \leq I(Y;Z) = \mu(Y \cap Z) = \mu(Y \cap Z \cap X) + \mu(Y \cap Z \cap X^c) \\
& = I(X;Y;Z) + I(Y;Z|X) = b + d \\
\end{align*}
\]

- Then we have

\[
\Rightarrow \\
I(X;Y|Z) = a \geq 0 \\
I(Y;Z|X) = b \geq 0 \\
H(Z|X,Y) = c \geq 0 \\
I(X;Y;Z) = d \\
\]

- Thus, \( d \geq -b \), and

\[
H(X) - H(Z) = (a + d) - (d + b + c) = a - b - c \\
\]
Imperfect Secrecy Theorem

Consider the next info diagram

\[ I(X;Y|Z) = a \geq 0 \]
\[ I(Y;Z|X) = b \geq 0 \]
\[ H(Z|X,Y) = c \geq 0 \]
\[ I(X;Y;Z) = d \]

Then we have

\[ 0 \leq I(Y;Z) = \mu(Y \cap Z) = \mu(Y \cap Z \cap X) + \mu(Y \cap Z \cap X^c) \]
\[ = I(X;Y;Z) + I(Y;Z|X) = b + d \]

Thus, \( d \geq -b \), and

\[ H(X) - H(Z) \]
Imperfect Secrecy Theorem

Consider the next info diagram

Then we have

\[ 0 \leq I(Y; Z) = \mu(Y \cap Z) = \mu(Y \cap Z \cap X) + \mu(Y \cap Z \cap X^c) \]
\[ = I(X; Y; Z) + I(Y; Z|X) = b + d \]  \hspace{1cm} (31.102)

Thus, \( d \geq -b \), and

\[ H(X) - H(Z) = (a + d) - (d + b + c) \]  \hspace{1cm} (31.103)
Imperfect Secrecy Theorem

Consider the next info diagram

\[ I(X; Y|Z) = a \geq 0 \]
\[ I(Y; Z|X) = b \geq 0 \]
\[ H(Z|X, Y) = c \geq 0 \]
\[ I(X; Y; Z) = d \]

Then we have

\[ 0 \leq I(Y; Z) = \mu(Y \cap Z) = \mu(Y \cap Z \cap X) + \mu(Y \cap Z \cap X^c) \]
\[ = I(X; Y; Z) + I(Y; Z|X) = b + d \]  \hspace{1cm} (31.101)

Thus, \( d \geq -b \), and

\[ H(X) - H(Z) = (a + d) - (d + b + c) = a - b - c \]  \hspace{1cm} (31.103)
Imperfect Secrecy Theorem

Consider the next info diagram

\[ I(X; Y|Z) = a \geq 0 \]
\[ I(Y; Z|X) = b \geq 0 \]
\[ H(Z|X, Y) = c \geq 0 \]
\[ I(X; Y; Z) = d \]

Then we have

\[ 0 \leq I(Y; Z) = \mu(Y \cap Z) = \mu(Y \cap Z \cap X) + \mu(Y \cap Z \cap X^c) \]

\[ = I(X; Y; Z) + I(Y; Z|X) = b + d \]

Thus, \( d \geq -b \), and

\[ H(X) - H(Z) = (a + d) - (d + b + c) = a - b - c \]

so \( I(X; Y) = a + d \geq a - b \geq a - b - c = H(X) - H(Z) \)
Note that a number of (nice) assumptions were not used in this derivation.
Imperfect Secrecy Theorem

Note that a number of (nice) assumptions were not used in this derivation.

1. Did not use $H(Y|X,Z) = 0$, so we could have random cipher text as a function of the text and password.
Imperfect Secrecy Theorem

- Note that a number of (nice) assumptions were not used in this derivation.

1. Did not use $H(Y|X, Z) = 0$, so we could have random cipher text as a function of the text and password.

2. $I(X; Z) = 0$, so don’t require that the text and password to be independent (so it can be made to be easy to remember the password even if it is long).
Imperfect Secrecy Theorem

- Note that a number of (nice) assumptions were not used in this derivation.

1. Did not use $H(Y|X, Z) = 0$, so we could have random cipher text as a function of the text and password.

2. $I(X; Z) = 0$, so don’t require that the text and password to be independent (so it can be made to be easy to remember the password even if it is long).

- But then, this is only a lower bound, it shows a possible worst case.
The Venn and Art of the Data Processing Inequality

- Markov chain $X_1 \rightarrow X_2 \rightarrow X_3 \rightarrow X_4$ gives:

Considering the above figure, we have:

$$\mu(\tilde{X}_1 \cap \tilde{X}_4) = I(X_1; X_4) \leq I(X_2; X_3) = \mu(\tilde{X}_2 \cap \tilde{X}_3)$$

(31.104)

We also have

$$\mu(\tilde{X}_2 \cap \tilde{X}_3) = I(X_2; X_3) \geq I(X_2; X_4) = \mu(\tilde{X}_2 \cap \tilde{X}_4)$$

(31.105)

And all of the others...
The Venn and Art of the Data Processing Inequality

- Markov chain $X_1 \to X_2 \to X_3 \to X_4$ gives:

$$X_1 \ X_2 \ X_3 \ X_4$$

- Considering the above figure, we have:

$$\mu(\tilde{X}_1 \cap \tilde{X}_4) = I(X_1; X_4) \leq I(X_2; X_3) = \mu(\tilde{X}_2 \cap \tilde{X}_3) \quad (31.104)$$

which is data processing inequality.
The Venn and Art of the Data Processing Inequality

- Markov chain $X_1 \rightarrow X_2 \rightarrow X_3 \rightarrow X_4$ gives:

  \[
  \mu(\tilde{X}_1 \cap \tilde{X}_4) = I(X_1; X_4) \leq I(X_2; X_3) = \mu(\tilde{X}_2 \cap \tilde{X}_3) \quad (31.104)
  \]

  which is data processing inequality.

- Considering the above figure, we have:

  \[
  \mu(\tilde{X}_2 \cap \tilde{X}_3) = I(X_2; X_3) \geq I(X_2; X_4) = \mu(\tilde{X}_2 \cap \tilde{X}_4) \quad (31.105)
  \]

  We also have
The Venn and Art of the Data Processing Inequality

- Markov chain $X_1 \rightarrow X_2 \rightarrow X_3 \rightarrow X_4$ gives:

  \[
  \mu(\tilde{X}_1 \cap \tilde{X}_4) = I(X_1; X_4) \leq I(X_2; X_3) = \mu(\tilde{X}_2 \cap \tilde{X}_3)
  \]  
  \[\text{ (31.104)}\]

  which is data processing inequality.

- Considering the above figure, we have:

  \[
  \mu(\tilde{X}_2 \cap \tilde{X}_3) = I(X_2; X_3) \geq I(X_2; X_4) = \mu(\tilde{X}_2 \cap \tilde{X}_4)
  \]  
  \[\text{ (31.105)}\]

- We also have

  \[
  \mu(\tilde{X}_1 \cap \tilde{X}_4) = I(X_1; X_4) \leq I(X_2; X_3) = \mu(\tilde{X}_2 \cap \tilde{X}_3)
  \]

- And all of the others . . .
Again, let $[n]$ be an index set of random variables and let $A, B, C \subseteq [n]$ and define $\alpha \triangleq A \cup C$ and $\beta \triangleq B \cup C$. Note that for conditional mutual information we have the following:

$$0 \leq I(X_A;X_B|X_C) = H(X_A|X_C) - H(X_A|X_B,X_C) \quad (31.106)$$

$$= H(X_A,X_C) - H(X_C) - H(X_A,B,C) + H(X_B,X_C) \quad (31.107)$$

$$= H(X_A \cup C) + H(X_B \cup C) - H(X_{\alpha \cap \beta}) - H(X_{\alpha \cup \beta}) \quad (31.108)$$

Thus, defining a function of the form $f(A) = H(X_A)$ this proves that for any $A, B \subseteq [n]$

$$f(A) + f(B) \geq f(A \cup B) + f(A \cap B) \quad (31.110)$$

meaning that the entropy function is submodular in the index set of random variables.
Again, let $[n]$ be an index set of random variables and let $A, B, C \subseteq [n]$ and define $\alpha \triangleq A \cup C$ and $\beta \triangleq B \cup C$.

Note that for conditional mutual information we have the following:

\[ 0 \leq I(X_A; X_B | X_C) = H(X_A | X_C) - H(X_A | X_B, X_C) \]
\[ = H(X_A, X_C) - H(X_C) - H(X_A, X_B, X_C) + H(X_B, X_C) \]
\[ = H(X_\alpha) + H(X_\beta) - H(X_{\alpha \cap \beta}) - H(X_{\alpha \cup \beta}) \]

(31.109)

Thus, defining a function of the form $f(A) = H(X_A)$ this proves that for any $A, B \subseteq [n]$

\[ f(A) + f(B) \geq f(A \cup B) + f(A \cap B) \]

(31.110)

meaning that the entropy function is submodular in the index set of random variables.
Again, let $[n]$ be an index set of random variables and let
$A, B, C \subseteq [n]$ and define $\alpha \triangleq A \cup C$ and $\beta \triangleq B \cup C$.

Note that for conditional mutual information we have the following:

$$I(X_A; X_B | X_C) = H(X_A | X_C) - H(X_A | X_B, X_C)$$

(31.106)

$$= H(X_A, X_C) - H(X_C) - H(X_A, X_B, X_C) + H(X_B, X_C)$$

(31.107)

$$= H(X_A \cup C) + H(X_B \cup C) - H(X_C) - H(X_A \cup B \cup C)$$

(31.108)

$$= H(X_\alpha) + H(X_\beta) - H(X_{\alpha \cap \beta}) - H(X_{\alpha \cup \beta})$$

(31.109)

Thus, defining a function of the form $f(A) = H(X_A)$ this proves
that for any $A, B \subseteq [n]$ 

$$f(A) + f(B) \geq f(A \cup B) + f(A \cap B)$$

(31.110)

meaning that the entropy function is
submodular

in the index set of
random variables.
Again, let \([n]\) be an index set of random variables and let \(A, B, C \subseteq [n]\) and define \(\alpha \triangleq A \cup C\) and \(\beta \triangleq B \cup C\).

Note that for conditional mutual information we have the following:

\[
0 \leq I(X_A; X_B | X_C)
\]

\[
= H(X_A | X_C) - H(X_A | X_B, X_C)
\]

\[
= H(X_A, X_C) - H(X_C) - H(X_A, X_B, X_C) + H(X_B, X_C)
\]

\[
= H(X_{\alpha}) + H(X_{\beta}) - H(X_{\alpha} \cap X_{\beta}) - H(X_{\alpha} \cup X_{\beta})
\]

Thus, defining a function of the form \(f(A) = H(X_A)\) this proves

\[
\text{that for any } A, B \subseteq [n] 
\]

\[
f(A) + f(B) \geq f(A \cup B) + f(A \cap B)
\]

meaning that the entropy function is submodular in the index set of random variables.
Again, let \([n]\) be an index set of random variables and let \(A, B, C \subseteq [n]\) and define \(\alpha \triangleq A \cup C\) and \(\beta \triangleq B \cup C\).

Note that for conditional mutual information we have the following:

\[
0 \leq I(X_A; X_B|X_C) = H(X_A|X_C) - H(X_A|X_B, X_C) \tag{31.106}
\]

Thus, defining a function of the form \(f(A) = H(X_A)\) this proves that for any \(A, B \subseteq [n]\)

\[
f(A) + f(B) \geq f(A \cup B) + f(A \cap B) \tag{31.110}
\]

meaning that the entropy function is submodular in the index set of random variables.
Again, let \([n]\) be an index set of random variables and let \(A, B, C \subseteq [n]\) and define \(\alpha \triangleq A \cup C\) and \(\beta \triangleq B \cup C\).

Note that for conditional mutual information we have the following:

\[
0 \leq I(X_A; X_B | X_C) = H(X_A | X_C) - H(X_A | X_B, X_C)
= H(X_A, X_C) - H(X_C) - H(X_A, X_B, X_C) + H(X_B, X_C)
\]  

(31.106) 

(31.107) 

Thus, defining a function of the form \(f(A) = H(X_A)\) this proves that for any \(A, B \subseteq [n]\)

\[
f(A) + f(B) \geq f(A \cup B) + f(A \cap B)
\]

(31.110) 

meaning that the entropy function is submodular in the index set of random variables.
Again, let \([n]\) be an index set of random variables and let \(A, B, C \subseteq [n]\) and define \(\alpha \triangleq A \cup C\) and \(\beta \triangleq B \cup C\).

Note that for conditional mutual information we have the following:

\[
0 \leq I(X_A; X_B|X_C) = H(X_A|X_C) - H(X_A|X_B, X_C) \tag{31.106}
\]

\[
= H(X_A, X_C) - H(X_C) - H(X_A, X_B, X_C) + H(X_B, X_C) \tag{31.107}
\]

\[
= H(X_{A \cup C}) + H(X_{B \cup C}) - H(X_C) - H(X_{A \cup B \cup C}) \tag{31.108}
\]

\[
\therefore f(A) + f(B) \geq f(A \cup B) + f(A \cap B) \tag{31.110}
\]

meaning that the entropy function is submodular in the index set of random variables.
Inequalities

- Again, let \([n]\) be an index set of random variables and let \(A, B, C \subseteq [n]\) and define \(\alpha \triangleq A \cup C\) and \(\beta \triangleq B \cup C\).
- Note that for conditional mutual information we have the following:

\[
0 \leq I(X_A; X_B | X_C) = H(X_A | X_C) - H(X_A | X_B, X_C) \tag{31.106}
\]

\[
= H(X_A, X_C) - H(X_C) - H(X_A, X_B, X_C) + H(X_B, X_C) \tag{31.107}
\]

\[
= H(X_{\alpha \cup C}) + H(X_{\beta \cup C}) - H(X_C) - H(X_{\alpha \cup B \cup C}) \tag{31.108}
\]

\[
= H(X_{\alpha \cup C}) + H(X_{\beta \cup C}) - H(X_{(A \cup C) \cap (B \cup C)}) - H(X_{(A \cup C) \cup (B \cup C)}) \tag{31.109}
\]
Again, let $[n]$ be an index set of random variables and let $A, B, C \subseteq [n]$ and define $\alpha \triangleq A \cup C$ and $\beta \triangleq B \cup C$.

Note that for conditional mutual information we have the following:

$$0 \leq I(X_A; X_B | X_C) = H(X_A | X_C) - H(X_A | X_B, X_C) \quad (31.106)$$

$$= H(X_A, X_C) - H(X_C) - H(X_A, X_B, X_C) + H(X_B, X_C) \quad (31.107)$$

$$= H(X_{A \cup C}) + H(X_{B \cup C}) - H(X_C) - H(X_{A \cup B \cup C}) \quad (31.108)$$

$$= H(X_{A \cup C}) + H(X_{B \cup C}) - H(X_{(A \cup C) \cap (B \cup C)}) - H(X_{(A \cup C) \cup (B \cup C)})$$

$$= H(X_\alpha) + H(X_\beta) - H(X_{\alpha \cap \beta}) - H(X_{\alpha \cup \beta}) \quad (31.109)$$
Again, let \([n]\) be an index set of random variables and let \(A, B, C \subseteq [n]\) and define \(\alpha \triangleq A \cup C\) and \(\beta \triangleq B \cup C\).

Note that for conditional mutual information we have the following:

\[
0 \leq I(X_A; X_B|X_C) = H(X_A|X_C) - H(X_A|X_B, X_C) = H(X_A, X_C) - H(X_C) - H(X_A, X_B, X_C) + H(X_B, X_C)
\]

\[
= H(X_{A \cup C}) + H(X_{B \cup C}) - H(X_C) - H(X_{A \cup B \cup C})\tag{31.108}
\]

\[
= H(X_{A \cup C}) + H(X_{B \cup C}) - H(X_{(A \cup C) \cap (B \cup C)}) - H(X_{(A \cup C) \cup (B \cup C)})
\]

\[
= H(X_{\alpha}) + H(X_{\beta}) - H(X_{\alpha \cap \beta}) - H(X_{\alpha \cup \beta})\tag{31.109}
\]

Thus, defining a function of the form \(f(A) = H(X_A)\) this proves that for any \(A, B \subseteq [n]\)

\[
f(A) + f(B) \geq f(A \cup B) + f(A \cap B)\tag{31.110}
\]

meaning that the entropy function is submodular in the index set of random variables.
We also know other properties of the entropy, namely

\[ H(X_\emptyset) = f(\emptyset) = 0 \quad (31.111) \]

and also that for any \( A \subseteq B \),

\[ f(A) = H(X_A) \leq H(X_B) = f(B) \quad (31.112) \]

meaning that entropy is a monotone non-decreasing function of the set of random variables (adding more random variables to discrete entropy function can only increase the uncertainty/information).
This begs the question: Consider all functions $f : 2^{[n]} \rightarrow \mathbb{R}_+$ that satisfy the following three properties:

$$f(\emptyset) = 0$$  \hspace{1cm} (31.113)

$$f(A) \leq f(B) \text{ whenever } A \subseteq B \subseteq [n]$$  \hspace{1cm} (31.114)

$$f(A) + f(B) \geq f(A \cup B) + f(A \cap B) \text{ for all } A, B \subseteq [n]$$  \hspace{1cm} (31.115)

and call this set of functions $\Gamma_n$ (note that this set is called the **cone of polymatroidal functions**, the reason being that they are closed under conic combinations).
This begs the question: Consider all functions $f : 2^{[n]} \rightarrow \mathbb{R}_+$ that satisfy the following three properties:

\begin{align*}
  f(\emptyset) &= 0 \quad (31.113) \\
  f(A) &\leq f(B) \text{ whenever } A \subseteq B \subseteq [n] \quad (31.114) \\
  f(A) + f(B) &\geq f(A \cup B) + f(A \cap B) \text{ for all } A, B \subseteq [n] \quad (31.115)
\end{align*}

and call this set of functions $\Gamma_n$ (note that this set is called the cone of polymatroidal functions, the reason being that they are closed under conic combinations).

Now, consider the set of all discrete entropy functions on $n$ random variables, call it $\Gamma^H_n$ (also close this set as well).
More inequalities

This begs the question: Consider all functions $f : 2^{[n]} \to \mathbb{R}_+$ that satisfy the following three properties:

$$f(\emptyset) = 0$$  \hspace{2cm} (31.113)

$$f(A) \leq f(B) \text{ whenever } A \subseteq B \subseteq [n]$$  \hspace{2cm} (31.114)

$$f(A) + f(B) \geq f(A \cup B) + f(A \cap B) \text{ for all } A, B \subseteq [n]$$  \hspace{2cm} (31.115)

and call this set of functions $\Gamma_n$ (note that this set is called the cone of polymatroidal functions, the reason being that they are closed under conic combinations).

Now, consider the set of all discrete entropy functions on $n$ random variables, call it $\Gamma^H_n$ (also close this set as well).

Is there a relationship between $\Gamma^H_n$ and $\Gamma_n$?
This begs the question: Consider all functions \( f : 2^n \rightarrow \mathbb{R}_+ \) that satisfy the following three properties:

\[
f(\emptyset) = 0 \quad (31.113)
\]

\[
f(A) \leq f(B) \text{ whenever } A \subseteq B \subseteq [n] \quad (31.114)
\]

\[
f(A) + f(B) \geq f(A \cup B) + f(A \cap B) \text{ for all } A, B \subseteq [n] \quad (31.115)
\]

and call this set of functions \( \Gamma_n \) (note that this set is called the cone of polymatroidal functions, the reason being that they are closed under conic combinations).

Now, consider the set of all discrete entropy functions on \( n \) random variables, call it \( \Gamma^H_n \) (also close this set as well).

Is there a relationship between \( \Gamma^H_n \) and \( \Gamma_n \)?

Clearly, \( \Gamma^H_n \subseteq \Gamma_n \) since entropy satisfies the three properties.
The following can be shown:
The following can be shown:

Theorem 31.3.4

*For* \( n = 2 \) or \( n = 3 \), we have that \( \Gamma^H_n = \Gamma_n \).
The following can be shown:

**Theorem 31.3.4**

*For* $n = 2$ or $n = 3$, we have that $\Gamma^H_n = \Gamma_n$.

So it may be the case that this would generalize for $n \geq 4$, right?
The following can be shown:

**Theorem 31.3.4**

For $n = 2$ or $n = 3$, we have that $\Gamma_{n}^{H} = \Gamma_{n}$.

- So it may be the case that this would generalize for $n \geq 4$, right?
- Would be a nice result as it would mean that entropy functions are as general as the set of polymatroidal functions (examples include, say, rank of a matrix, matroid rank, grouped matroid rank, weighted matroid rank, and many others).
Non-Shannon Information Theory Inequality

Via information measures ideas presented earlier, we can get:

**Theorem 31.3.5**

For any four discrete random variables $X, Y, Z, U$, define

$$
\Delta(Z, U|X, Y) = I(Z; U) - I(Z; U, X) - I(Z; U, Y)
$$

then the following inequality holds:

$$
\Delta(Z, U|X, Y) \leq \frac{1}{2} \left[ I(X; Y) + I(X; Z, U) + I(Z; U, X) - I(Z; U|Y) \right]
$$

(31.117)

and while the l.h.s. is symmetric in $X$ and $Y$ but the r.h.s. is not, we also have:

$$
\Delta(Z, U|X, Y) \leq \frac{1}{2} \left[ I(X; Y) + I(Y; Z, U) - I(Z; U, X) + I(Z; U|Y) \right]
$$

(31.118)
Consequences of this Inequality

Define $\Delta(i, j|k, \ell) \triangleq I(\{i\}; \{j\}) - I(\{i\}; \{j\}, \{k\}) - I(\{i\}; \{j\}, \{\ell\})$
where whenever we mention $\{i\}$ we really mean the r.v. $X_{\{i\}}$. 

Define $\tilde{\Gamma}_4$ as follows:
$\tilde{\Gamma}_4 = \{ F \in \Gamma_4 : \text{for all permutations } \sigma \text{ of } \{1, 2, 3, 4\}, \Delta(\sigma(1), \sigma(2)|\sigma(3), \sigma(4)) \leq \frac{1}{2} [I(\sigma(3); \sigma(4)) + I(\sigma(1); \sigma(2), \sigma(3)) - I(\sigma(1); \sigma(2), \sigma(4)) + I(\sigma(3); \sigma(1), \sigma(2))] \}$
So, $\tilde{\Gamma}_4$ are the polymatroidal functions that also satisfy the additional information theoretic inequality and so are a subclass that includes the entropy functions.
Consequences of this Inequality

- Define $\Delta(i, j|k, \ell) \triangleq I(\{i\}; \{j\}) - I(\{i\}; \{j\}, \{k\}) - I(\{i\}; \{j\}, \{\ell\})$ where whenever we mention $\{i\}$ we really mean the r.v. $X_{\{i\}}$.

- Define $\tilde{\Gamma}_4$ as follows:

$$\tilde{\Gamma}_4 = \left\{ F \in \Gamma_4 : \text{for all permutations } \sigma \text{ of } \{1, 2, 3, 4\}, \right.$$ 
$$\Delta(\sigma(1), \sigma(2)|\sigma(3), \sigma(4)) \leq \frac{1}{2} \left[ I(\sigma(3); \sigma(4)) + I(\sigma(1); \sigma(2), \sigma(3)) \right.$$ 
$$- I(\sigma(1); \sigma(2), \sigma(4)) + I(\sigma(3); \sigma(1), \sigma(2)) \right\}$$
Consequences of this Inequality

- Define $\Delta(i, j|k, \ell) \triangleq I(\{i\}; \{j\}) - I(\{i\}; \{j\}, \{k\}) - I(\{i\}; \{j\}, \{\ell\})$ where whenever we mention $\{i\}$ we really mean the r.v. $X_{\{i\}}$.

- Define $\tilde{\Gamma}_4$ as follows:

$$\tilde{\Gamma}_4 = \left\{ F \in \Gamma_4 : \text{for all permutations } \sigma \text{ of } \{1, 2, 3, 4\}, \right.$$ 

$$\Delta(\sigma(1), \sigma(2)|\sigma(3), \sigma(4)) \leq \frac{1}{2} \left[ I(\sigma(3); \sigma(4)) + I(\sigma(1); \sigma(2), \sigma(3)) - I(\sigma(1); \sigma(2), \sigma(4)) + I(\sigma(3); \sigma(1), \sigma(2)) \right] \right\}$$

- So, $\tilde{\Gamma}_4$ are the polymatroidal functions that also satisfy the additional information theoretic inequality and so are a subclass that includes the entropy functions.
Theorem 31.3.6

\[ \Gamma_n^H \neq \Gamma_n \] (31.119)

Proof.
Consequences of this Inequality

Theorem 31.3.6

\[
\Gamma_n^H \neq \Gamma_n
\]  

(31.119)

Proof.

Define a function \( F \) as follows:

\[
F(\emptyset) = 0,
F(X) = F(Y) = F(Z) = F(U) = 2a > 0,
F(X, Y) = 4a
\]

\[
F(X, U) = F(X, Z) = F(Y, U) = F(Y, Z) = F(Z, U) = 3a
\]

\[
F(X, Y, Z) = F(X, Y, U) = F(X, Z, U)
= F(Y, Z, U) = F(X, Y, Z, U) = 4a
\]
Info Measures

**Consequences of this Inequality**

**Proof.**

- This function is polymatroidal (i.e., $F \in \Gamma_4$) but it is not in the class that contains the entropy function (i.e., $F \notin \tilde{\Gamma}_4$) (exercise: verify this).
Consequences of this Inequality

Proof.

- This function is polymatroidal (i.e., $F \in \Gamma_4$) but it is not in the class that contains the entropy function (i.e., $F \notin \widetilde{\Gamma}_4$) (exercise: verify this)

- Therefore, the polymatroidal class is strictly greater than the class of entropy functions.